Universitat d'Alacant
Universidad de Alicante

ESTADÍSTICA AVANZADA EN CIENCIAS DE LA SALUD:

Modelos Lineales

Diciembre de 2016

Introducción

Este material presenta métodos estadísticos avanzados para el entorno de las ciencias de la salud. Proviene de la experiencia docente de su autor, profesor en asignaturas, cursos y seminarios impartidos en el entorno de las ciencias de la salud, dirigidos a profesionales que desarrollan su labor asistencial, de gestión, de docencia o de investigación en este ámbito.

Los contenidos aquí recogidos tratan de los modelos multivariantes de carácter probabilístico más frecuentes en el ámbito de la investigación en estudios observacionales en el entorno de las Ciencias de la Salud. Dan respuesta a las preguntas de investigación clínico-epidemiológica que plantean el análisis de si una o más variables (explicativas) aportan capacidad explicativa relevante sobre una variable resultado (respuesta) bajo una estructura lineal. Los modelos lineales aquí recogidos permiten utilizar variables respuesta de tipo cuantitativo (regresión lineal múltiple), o basadas en variables de tipo dicotómico, a través de modelizar la incidencia acumulada de un resultado (regresión logística), la tasa o densidad de indicdencia del resultado (regresión de Poisson) o la tasa instantánea de incidencia en el tiempo del resultado (regresión de Cox).

Los contenidos de este texto están especialmente orientados como material de apoyo teórico-práctico en cursos, seminarios, etc. cuyo objetivo sea el de profundizar en estos modelos. Estos materiales han venido mostrando su utilidad como soporte docente en asignaturas impartidas en las titulaciones de máster de Ciencias de la Salud de la Universidad de Alicante, como Investigación de Ciencias de la Salud, Salud Pública, Óptica,.., y en numerosos cursos de posgrado impartidos por el autor.

El material se estructura en cinco capítulos, partiendo de un capítulo introductorio en el que se exponen las características generales del análisis multivariante y las propiedades y características comunes de los modelos que se desarrollan. A continuación se presenta un capítulo dedicado a cada uno de los modelos. La presentación de los modelos sigue el esquema estructura, construcción-estimación, requerimientos e inferencias con el modelo. El material se presenta en un formato de dos columnas, la primera de ellas (izquierda) está dedicada a presentar los contenidos teóricos de apoyo, mientras que la segunda (derecha) se dedica a aplicaciones de revisión-ejemplificación para que el alumno aplique los contenidos teóricos.

Alicante, otoño de 2016
El autor

Andreu Nolasco Bonmatí es profesor del Departamento de Enfermería Comunitaria, Medicina Preventiva y Salud Pública e Historia de la Ciencia de la Universidad de Alicante en el que ha venido desarrollando su labor como docente e investigador. Ha impartido docencia en diversas titulaciones de Ciencias de la Salud (Medicina, Enfermería, Nutrición humana y dietética, Óptica, etc.) tanto en estudios de grado como en posgrado (máster y doctorado), en materias y/o asignaturas como Bioestadística, Estadística Avanzada, Demografía y Salud, Metodología de la Investigación, Desigualdades en Salud, Análisis de la mortalidad, etc.. Ha venido desarrollando investigación en líneas como: Análisis de la Mortalidad, Geografía Sanitaria, Estadísticas Sanitarias, Encuestas de salud, Demografía y salud, Desigualdades en salud y otras. Su experiencia en la aplicación del método estadístico en el entorno de las Ciencias de la Salud proviene y se refleja en la dirección de numerosos proyectos de investigación, tesis doctorales, publicaciones científicas y el continuo contacto con el contexto sanitario a través del asesoramiento metodológico a diversas instituciones sanitarias (Administración sanitaria, Centros de Salud y Salud Pública, Hospitales y otras).

SUMARIO

Introducción al análisis multivariante. Los modelos lineales.....6
El modelo de regresión lineal múltiple 16
El modelo de regresión logística.....39
El modelo de regresión de Poisson.....67
El modelo de regresión de Cox.....81
Bibliografía.....95

INTRODUCCION AL ANALISIS MULTIVARIANTE. LOS MODELOS LINEALES

Una definición de análisis multivariante

Conjunto de técnicas estadísticas basadas en el estudio conjunto de varias variables (3 o más) con el objetivo de describir o hacer inferencias sobre las características individuales o colectivas de tales variables

Sugiera alguna situación de análisis multivariante

P.ej. Estudio de prevalencia de cálculos biliares en población general y su relación con el sexo, la edad y el consumo de cítricos

Una reflexión

Con esta definición el análisis multivariante abarcará la mayor parte de las aplicaciones estadísticas

¿Podemos construir una clasificación de los métodos multivariantes?

Diversos autores proponen clasificaciones diferentes de los métodos multivariantes:

Enumere aquellas técnicas de análisis estadístico univariante que recuerde

P.ej. Estimación por intervalos de confianza de un parámetro desconocido

UNA CLASIFICACION ESENCIALMENTE TEORICA	
METODOS MULTIVARIANTES NO PROBABILISTICOS	METODOS MULTIVARIANTES PROBABILISTICOS
Se fundamentan en resultados no probabilísticos. No suponen distribuciones de probabilidad subyacentes. Utilizan resultados basados en álgebra lineal (geométricos). Tienen poca capacidad para producir inferencias	Se fundamentan en resultados derivados de la teoría de la probabilidad, suposiciones probabilísticas sobre las variables a estudio y/o análisis de la función de verosimilitud. Son idóneos para producir inferencias

UNA CLASIFICACION ESENCIALMENTE OPERATIVA	
METODOS MULTIVARIANTES QUE CONSIDERAN A TODAS LAS VARIABLES EN EL MISMO "STATUS"	METODOS MULTIVARIANTES QUE PARTEN DE DOS CONJUNTOS DE VARIABLES: EXPLICATIVAS Y RESPUESTA
En origen no asignan un tratamiento diferente a unas u otras variables. Generalmente estos métodos coinciden con la definición no probabilística	Parten de que las variables involucradas pertenecen de forma clara (o moderadamente clara) a dos conjuntos, las que se ven influidas (**variables respuesta**) y las que influyen (**variables explicativas**). Coinciden en su mayor parte con la definición probabilística

¿Se le ocurre alguna situación en la que pudiéramos aplicar métodos no probabilísticos? Coméntela

Un pequeño resumen de métodos multivariantes no probabilísticos

Forman parte destacada de estos métodos los llamados métodos de estadística descriptiva multidimensional:

METODOS FACTORIALES	METODOS DE CLASIFICACION
- **Análisis factorial clásico** - **Análisis de componentes principales** - **Análisis de correspondencias** - **Análisis canónico, de los rangos**	- **Cluster jerárquico** - **Cluster no jerárquico** - **Análisis discriminante**

Métodos multivariantes probabilísticos: La tipología de las variables a estudio

Clasifique las variables del estudio descrito según su tipo

Para introducir estas ideas, considere como ejemplo un estudio en el que las variables a estudio son:

Tabla 1

COLESTOT	= Nivel de colesterol medido en mg/100ml
QUETELET	= Indice de quetelet en sus unidades
EDAD	= Edad en años
ALCOHOL	= Consumo de alcohol: 0 'nunca' 1 'bajo' 2 'moderado/alto'

Suponga que el investigador persigue averiguar si el consumo de alcohol es un factor de riesgo con efecto sobre el colesterol pero teniendo en cuenta la edad de los individuos.

- **Clasificación por el tipo de variables:**

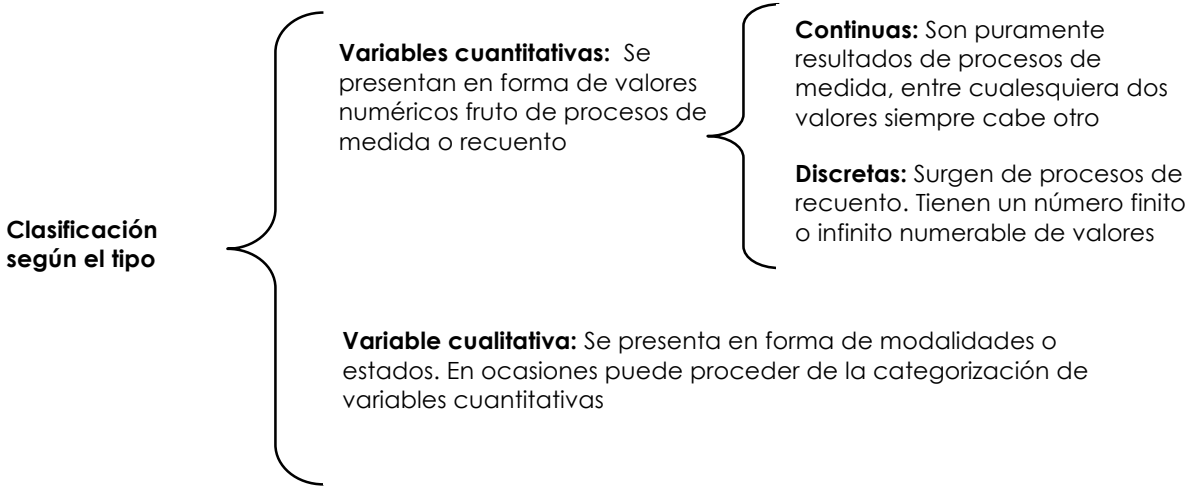


Figura 1.- Clasificación según el tipo de variables

- **Clasificación por el papel de las variables:**

Clasificación según el papel en el análisis

Variable respuesta: Aquella sobre la que se produce el efecto de las demás. Sus variaciones son consecuencia de las variaciones de las demás. También es denominada variable dependiente

Variables explicativas: El resto de variables. Se trata de determinar en que forma afectan a la respuesta. También se llaman independientes o predictores

- **Factores:** El interés del estudio estriba en detectar caracterizar su relación y efecto con la variable respuesta.
- **Covariables:** Son variables incluidas en el análisis por su posible relación con la respuesta y los factores.

Figura 2.- Clasificación de variables según su papel en el análisis

Clasificación de las covariables (variables de control)

De confusión: La relación entre el factor y la respuesta es diferente de tener en cuenta en el análisis a la covariable a no hacerlo. Para que una covariable sea de confusión es necesario y suficiente que se relacione con ambas, el factor y la respuesta. Debe tenerse en cuenta que lo que se confunde es la relación entre factor y respuesta

De interacción: La relación entre el factor y la respuesta se modifica según los valores o niveles de la covariable. Diremos que el efecto y/o relación entre factor y variable respuesta es inseparable de la covariable.

Figura 3.- Clasificación de las covariables

Teniendo en cuenta el objetivo del investigador, clasifique las variables del estudio descrito según su papel en el análisis

Si hay covariables, clasifíquelas

Los modelos estadísticos ¿Qué es un modelo estadístico?

Es imposible hablar de técnicas multivariantes sin recurrir al concepto de modelo estadístico.

Podríamos decir que un modelo estadístico no es más que una ecuación matemática con la que intentamos representar lo más fielmente posible la realidad a partir de la información que nos provee un conjunto de datos.

Pero conviene matizar o situar esta definición. Así, conviene pensar en un modelo estadístico como el resultado de la composición de dos elementos:

Componente estructural (sistemática): Parte determinista del modelo. Con ella **nosotros** asignamos *a priori* cual es la estructura esperada de la realidad. Generalmente es una **ecuación**

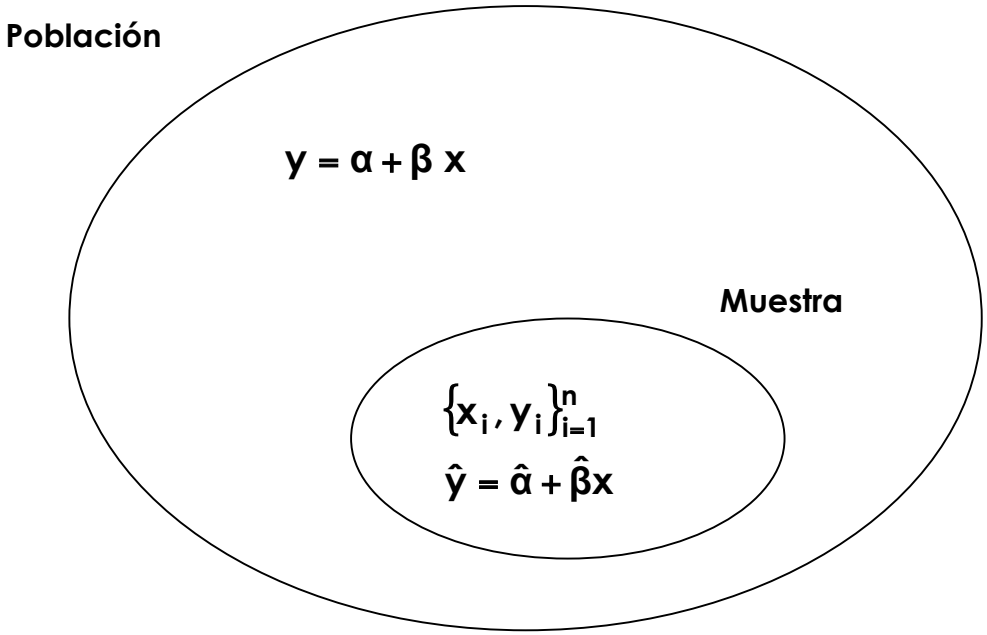
Componente aleatoria (estocástica): Parte en la que interviene el **azar** a través de los datos de una **muestra aleatoria.** Con esta componente hemos de evaluar la bondad de la estructura propuesta

Suponga que queremos modelizar la posible relación existente entre la proporción de piezas con caries y la edad en niños (menores de 16 años). Proponga componentes estructural y aleatoria para su construcción y discusión

El marco para las componentes estructural y aleatoria

Sean y = Proporción de caries x = Edad

El marco para las componentes estructural y aleatoria será:

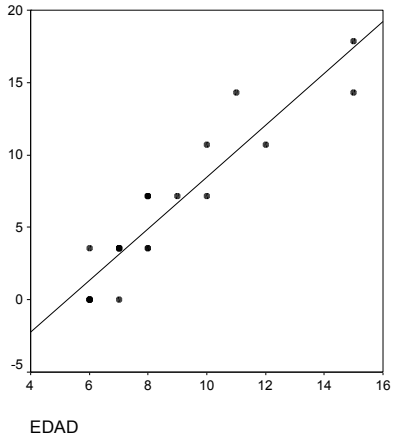


$y = \alpha + \beta x$ Modelo propuesto $\hat{y} = \hat{\alpha} + \hat{\beta}x$ Modelo estimado a

partir de $\{x_i, y_i\}_{i=1}^n$ Muestra aleatoria de n observaciones de x e y

α , β Parámetros desconocidos $\hat{\alpha}$, $\hat{\beta}$ Estadísticos muestrales conocidos

El gráfico adjunto muestra el diagrama de dispersión y recta de regresión sobre una muestra de n=24 niños



Ecuación estimada:

$$\hat{y} = -9{,}8 + 1{,}8x$$

¿Cómo sería un gráfico con:

- **Mucha evidencia de que la relación no es lineal habiéndola supuesto lineal**
- **Poca evidencia de una relación lineal que realmente no lo es**

Discuta estas cuestiones

Pero, ¿Cómo es la realidad?

Desde la perspectiva estadística es muy poco probable que la realidad responda a un único modelo. Quizás conviene partir desde la creencia de que la naturaleza es básicamente simple, es decir, lineal, siempre que puede.

Aunque algunos autores proponen lo contrario, la elección de una u otra forma estructural para nuestro modelo puede venir guiada por aspectos tales como:

- El tipo de variables
- El rango de las variables
- La verosimilitud de los datos
- La oportuna interpretación de los parámetros del modelo (que nos digan algo sencillo de entender)
- El objetivo del estudio (detección de asociaciones/relaciones, predicciones, ambos, etc...)

Estos aspectos pueden ayudarnos a decidir cual puede ser el modelo propuesto (su estructura). Así, tras verificar la bondad de tal modelo nos encontraremos en una de las dos situaciones siguientes:

El modelo no es aceptable

En este caso debemos pensar en que las variables estudiadas no se relacionan bajo el modelo propuesto. Quizás lo hagan bajo otro modelo

El modelo es aceptable

Las variables se relacionan bajo el modelo propuesto, pero nada garantiza que ese sea ni el único ni el mejor modelo para relacionarlas

Sugiera algunas situaciones para las que conozca que existe relación más o menos demostrada entre las variables implicadas. Si lo sabe, diga el modelo bajo el que se relacionan

(P.ej., peso y talla, modelo lineal)

Los modelos lineales generalizados (GLM)

Familia de modelos cuya componente estructural es:

$$f(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_i x_i + \cdots + \beta_k x_k$$

donde:

y = Variable respuesta. Aquella variable sobre la que deseamos medir el efecto de otras variables (tasa de mortalidad, probabilidad de efermar, nivel de ácido úrico,...)

$x = (x_1, x_2, \ldots, x_i, \ldots, x_k)$ = Variables explicativas. Variables incluidas en el modelo que suponemos que tienen capacidad para explicar las modificaciones que se producen en la variable respuesta. Pueden ser factores de riesgo o covariables

f(y) = **función nexo.** Es una función a través de la cual proponemos que se relaciona la variable respuesta con las explicativas

$(\beta_0, \beta_1, \beta_2, \ldots, \beta_i, \ldots, \beta_k)$ = Parámetros del modelo. Son coeficientes desconocidos de cuya estimación e inferencias obtendremos una cuantificación de las interrelaciones entre las variables explicativas y la variable respuesta

Algunos modelos de la clase GLM

La siguiente tabla muestra las características más destacables de algunos modelo de esta clase (* indica los que se verán en este curso)

Denominación Modelo	Variable respuesta (y)	Función nexo (f(y))
Regresión lineal Múltiple*	**y = variable cuantitativa** $-\infty < y < \infty$	**Identidad** $f(y) = y$
Regresión logística binaria*	**y = probabilidad** $0 < y < 1$	**Función logística** $f(y) = \log \frac{y}{1-y}$
Modelo log-lineal	**y = frecuencia** $0 < y$	**Logaritmo** $f(y) = \log y$
Regresión de Poisson*	**y = tasa media (densidad de incidencia)** $0 < y$	**Logaritmo** $f(y) = \log y$
Regresión de Cox*	**y(t) = tasa instantánea, varía en el tiempo** $0 < y$	**Logaritmo** $f_t(y) = \log y(t)$

Describa alguna situación de análisis en las que la variable de interés sea como las variables respuesta de los modelos de la tabla

Características más importantes de los modelos GLM

Revise la idea de aditividad. Contrástela frente a la idea de multiplicación de efectos

El efecto de las variables explicativas es supuesto aditivo sobre la función nexo. En efecto, las **(x_1, x_2, ..., x_i,...,x_k)** suman su efecto para producir cambios en la función nexo. En general, este efecto evaluado sobre la variable respuesta dependerá de la función nexo. Se verá para cada modelo particular

Demuestre la interpretación de los parámetros del modelo. ¿Qué interpretación tiene β_0

Los parámetros del modelo miden el cambio (incremento) de la función nexo por unidad de cambio en cada variable explicativa, manteniendo constante el resto de variables (ajustado por el resto de variables explicativas). La interpretación sobre la variable respuesta debe ser deducida para cada modelo a través de la función nexo

La mecánica en el proceso de estimación de modelos, inferencias o selección de modelos óptimos es semejante para todos ellos. Aunque variarán los elementos concretos de unos a otros modelos, los principios de evaluación de los modelos son los mismos. Comparten criterios de estimación de parámetros

El MODELO DE REGRESION LINEAL MULTIPLE

Modelo GLM con función nexo la identidad. Este modelo responde a la pregunta ¿de qué forma afectan k variables explicativas (cualitativas o cuantitativas) a una variable respuesta cuantitativa?. Debemos tener en cuenta que el objetivo puede ser:

- Detectar que variables, de entre las k consideradas, afectan o explican a la variable respuesta (detectar asociaciones o relaciones) y cuantificar el efecto

- Construir un modelo para predecir el valor de la variable respuesta en función de las variables explicativas

- Ambos

• Debemos tener en cuenta que aunque las variables explicativas estudiadas afecten el comportamiento de la respuesta, probablemente existan otras variables no incluidas entre ellas que también tengan efecto sobre la respuesta. La razón para no incluirlas en el modelo puede provenir de que no las conocemos o de que no queremos incluirlas.

EL INVESTIGADOR ES QUIEN DECIDE CUALES SON LAS VARIABLES A ESTUDIO (EXPLICATIVAS Y RESPUESTA) → Carga de subjetividad en la estructura propuesta

***Ejemplo 1.*- Suponga un estudio en el que el que se dispone de las variables descritas en la tabla 1. Sugiera algunas preguntas de interés que pudieran ser contestadas con un modelo de regresión lineal múltiple.**

¿Se le ocurre alguna variable no incluida entre las estudiadas que pudiera tener efecto sobre el nivel de colesterol?

Estructura del modelo de regresión

Exprese en términos de modelos de regresión las preguntas sugeridas en el ejemplo 1

La estructura del modelo es:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_i x_i + \cdots + \beta_k x_k + u$$

y = Variable respuesta
x = (x_1, x_2, ..., x_i,...,x_k) = Vector de variables explicativas
(β_0, β_1, β_2,...,β_i,...,β_k) = Parámetros del modelo
u = f(x_{k+1}, ...,x_h)= Perturbación o error, función de otras variables no incluidas en el modelo

Requerimientos/Hipótesis del modelo de regresión

La utilización inferencial plena del modelo de regresión requiere:

1. Para cada conjunto fijo de x la distribución de y debe ser normal con media
$$E(y/x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_i x_i + \cdots + \beta_k x_k$$
2. La varianza de y es constante para cualquier valor de x
3. Las observaciones de y son independientes entre sí
4. El número de variables explicativas es menor que el de observaciones

Comentarios a los requerimientos

Reflexione y discuta acerca de los requerimientos del modelo de regresión. ¿Cree que estos requerimientos serían aceptables sobre el ejemplo 1?

1. El requerimiento 1 establece que la ecuación lineal es pertinente. El requerimiento de normalidad tiene fundamentalmente dos objetivos:

   - Hacer coincidir el tipo de estimación mínimo cuadrática con la estimación máximo verosímil
   - Servir de base para las inferencias. Es el soporte probabilístico del modelo

2. El requerimiento 2 se denomina homocedasticidad. Su efecto se produce sobre la realización de inferencias

3. El requerimiento 3 establece que el conocimiento de unos valores de y no proporciona información para el conocimiento de otros valores

4. Requerimiento necesario para poder estimar el modelo sin ambiguedades

5. Los requerimientos enunciados suelen expresarse sintéticamente en términos de requerimientos sobre los errores u:

   - Los errores u siguen una distribución:

$$u \approx \text{Normal}(0, \sigma^2)$$

   - Los errores u son independientes entre sí (ausencia de autocorrelación)

Comprobación de los requerimientos del modelo

- La pertinencia del modelo lineal será comprobada a través de pruebas de bondad de ajuste

- Las hipótesis de normalidad, homocedasticidad e independencia se resolverán a través del análisis de los errores o residuos

Construcción de un modelo de regresión lineal múltiple. Etapas

Suponga que sobre una muestra de 200 sujetos seleccionados aleatoriamente de cierta población se dispone de las variables ya descritas en la tabla 1:

COLESTOT	= Nivel de colesterol medido en mg/100ml
QUETELET	= Indice de quetelet en sus unidades
EDAD	= Edad en años
ALCOHOL	= Consumo de alcohol: 0 'nunca' 1 'bajo' 2 'moderado/alto'

Puede consultar los datos de este ejemplo en el Anexo 1.

<u>Etapa 1</u>: Especificación de variables y modelo propuesto

Se desea averiguar si las variables Quetelet y Edad explican el Colestot. Si denotamos por

y = Colestot
x_1 = Quetelet
x_2 = Edad

Modelo propuesto

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Notas

Etapa 2: Estimación del modelo

A partir de los datos disponibles, una muestra aleatoria de n observaciones de las variabes:

$$\{y_i, x_{1i}, x_{2i}, \cdots, x_{ki}\}_{i=1}^{n}$$

Se trata de calcular los estimadores muestrales de los parámetros del modelo:

$(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k)$ estimadores de los parámetros $(\beta_0, \beta_1, \beta_2, \ldots, \beta_k)$

El método utilizado para ello suele ser el de mínimos cuadrados. Este método se basa en minimizar la función:

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

donde $$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \cdots + \hat{\beta}_k x_{ki}$$

Si los requerimientos del modelo se cumplen, el teorema de Gauss-Markov garantiza que los estimadores obtenidos son insesgados, óptimos (de mínima varianza) y coinciden con los obtenidos por máxima verosimilitud. Se obtiene así el modelo estimado:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k$$

La aplicación de este método al ejemplo sugerido da como resultado el modelo estimado:

Interprete los parámetros del modelo construido entre colesterol, edad y quetelet

Coeficientes[a]

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.
		B	Error típ.	Beta		
1	(Constante)	108,041	23,926		4,516	,000
	EDAD	,989	,233	,301	4,247	,000
	QUETELET	2,704	,971	,197	2,784	,006

a. Variable dependiente: COLESTOT

$$\hat{y} = 108{,}041 + 0{,}989\ \text{Edad} + 2{,}704\ \text{Quetelet}$$

Etapa 3.- Validación de la hipótesis de linealidad. Bondad de ajuste del modelo

• La evaluación de la bondad de ajuste del modelo puede hacerse a través de las variabilidades asociadas al modelo propuesto:

$\sum_{i=1}^{n} (y_i - \bar{y})^2$ Variabilidad de la variable respuesta. No depende del modelo

$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ Variabilidad del error. Es la parte no explicada por el modelo

$\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$ Variabilidad de la regresión. Es la explicada por el modelo

Puede demostrarse sin dificultad que:

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Profundice en el resultado R^2=0,178 y su interpretación

Podemos cuantificar la bondad de ajuste del modelo (capacidad para explicar la variabilidad de la variable respuesta a través de las variables consideradas y el modelo lineal supuesto) a través de:

Coeficiente de determinación

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\text{Variabilidad explicada}}{\text{Variabilidad total}} \qquad (0 \le R^2 \le 1)$$

R^2 es el estimador muestral del coeficiente de determinación poblacional, ρ^2.

Sobre el ejemplo

$$R^2 = 0{,}178$$

Este resultado nos indica que la capacidad explicativa (a través del modelo linela) de la edad y el índice de quetelet sobre el nivel de colesterol en los datos de la muestra es de 0,178 (17,8% de la variabilidad total del colesterol)

• La generalización inferencial de la bondad de ajuste del modelo puede realizarse a través de la prueba de hipótesis:

$$H_0 : \rho^2 = 0 \qquad H_a : \rho^2 \neq 0$$

con la que contrastaremos si las variables carecen de capacidad explicativa a través del modelo lineal en la población (H_0) frente al resultado de que tienen alguna capacidad explicativa a través del modelo lineal en la población (H_a). La solución al contraste se obtiene a través de la tabla del análisis de la variabilidad (ANOVA) de la regresión:

Fuente de variabilidad	Suma de cuadrados	Grados de libertad (gl)	Cuadrado medio	Estadístico de contraste
Regresión	$\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$	k	$\hat{s}_r^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{k}$	$F = \frac{\hat{s}_r^2}{\hat{s}_e^2}$
Error o residual	$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	n-k-1	$\hat{s}_e^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-k-1}$	Sigue una F de Snedecor con k y n-k-1 grados de libertad
Total	$\sum_{i=1}^{n}(y_i - \bar{y})^2$	n-1	$\hat{s}_y^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$	

Profundice en la interpretación de las hipótesis del contraste de bondad del modelo. Proponga algún enunciado equivalente.

Sobre el ejemplo se tiene:

Profundice en la interpretación del resultado del contraste de hipótesis de la bondad del modelo.

ANOVA[b]

Modelo		Suma de cuadrados	gl	Media cuadrática	F	Sig.
1	Regresión	90761,468	2	45380,734	21,270	,000[a]
	Residual	418174,240	196	2133,542		
	Total	508935,709	198			

a. Variables predictoras: (Constante), QUETELET, EDAD

b. Variable dependiente: COLESTOT

Resultado que conduce al rechazo de la hipótesis nula

Etapa 4.- Verificación de requerimientos. Análisis de residuos

Una vez estimado el modelo, los requerimientos del modelo pueden ser discutidos a través de los residuos:

$$e_i = y_i - \hat{y}_i$$

o sus transformaciones:

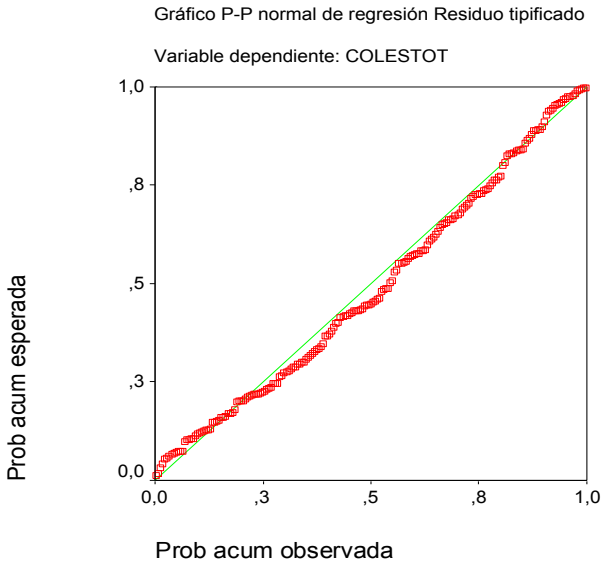
$$z_i = \frac{e_i}{\sqrt{Var(e_i)}} = \frac{e_i}{S} \qquad r_i = \frac{e_i}{S\sqrt{1-h_i}} \qquad r_{-i} = \frac{e_i}{S_{-i}\sqrt{1-h_i}}$$

Residuo estandarizado Residuo *estudentizado* Residuo por eliminación

h_i= elemento i de la diagonal de $X(X'X)^{-1}X'$, S_{-i}=Desviación estándar residual calculada cuando la observación i es eliminada

Si los requerimientos del modelo se cumplen los residuos estandarizados estudentizados y por eliminación siguen una distribución t de student con n-k-1 gl (estudentizados) y n-k-2 gl (por eliminación). A poco que n sea grande, n≥30, su distribución puede aproximarse por una normal. En general los residuos por eliminación suelen ser más utilizados por sus mejores propiedades de comportamiento distribucional.

• La normalidad de los residuos puede ser comprobada graficamente y/o a través de una prueba de normalidad:



Prueba de Kolmogorov-Smirnov para una muestra

		ZRE_1 Standardized Residual
N		199
Parámetros normales[a,b]	Media	,0000000
	Desviación típica	,99493670
Diferencias más extremas	Absoluta	,060
	Positiva	,060
	Negativa	-,032
Z de Kolmogorov-Smirnov		,849
Sig. asintót. (bilateral)		,467

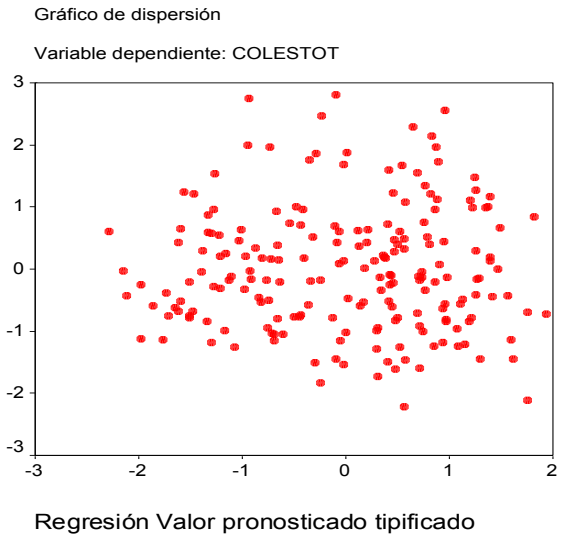
a. La distribución de contraste es la Normal.

b. Se han calculado a partir de los datos.

Interprete el gráfico y tabla adjuntas

Interprete el gráfico y el valor del estadístico de Durbin Watson

- La homocedasticidad u homogeneidad de varianzas puede ser discutida aproximadamente mediante el gráfico de los residuos frente a los valores estimados:



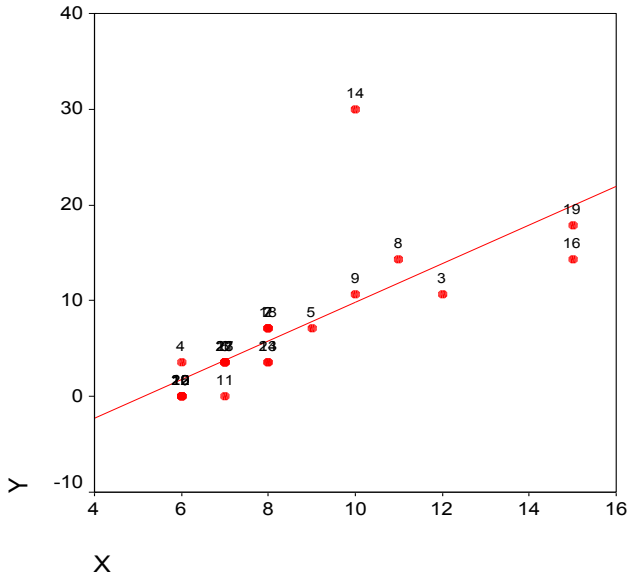
El gráfico no debe presentar distribución en forma de embudos, debiendo ser más o menos homogéneo.

- La independencia o ausencia de autocorrelación se puede discutir con el estadístico de Durbin-Watson, que contrasta la hipótesis nula de ausencia de autocorrelación (independencia) frente a lo contrario:

Durbin-Watson = 2,007

El valor debe ser contrastado en las tablas específicas para este estadístico. Un valor cercano a 2 es indicativo de aceptación de la hipótesis nula

• Los residuos pueden ser útiles para la detección de valores atípicos (*outliers*), es decir, indiciduos tales que una vez establecido el modelo,se apartan de la tendencia general del modelo. Como ejemplo observe el gráfico adjunto:



en el que puede observarse como el individuo 14 presenta un valor de y 'atípico' dado su valor de x en el modelo.

Si los residuos son normales, un test aproximado para la detección de estos valores se puede obtener a partir de los residuos estandarizados que seguirán una normal de media 0 y desviación típica 1. Así , por ejemplo, si establecemos colas de probabilidad 0,05, los valores superiores a 1,96 o inferiores a –1,96 tendrán una probabilidad inferior o igual a 0,05. El criterio de 'rareza' debe establecerlo el investigador (debe aplicarse Bonferroni). Este criterio suele ser utilizado preferentemente con residuos estudentizados o por eliminación

El listado adjunto incluye los casos para los que el residuo estandarizado es superior a 2 o inferior a –2. Discuta este criterio para los valores atípicos. Revise los casos atípicos en el anexo 1

Diagnósticos por caso[a]

Número de caso	Residuo tip.	COLESTOT	Valor pronosticado
14	2,142	344	245,07
25	2,291	347	241,19
41	-2,116	167	264,76
49	2,464	336	222,17
88	2,810	355	225,22
92	2,743	334	207,28
130	2,562	366	247,65
140	-2,213	137	239,23

a. Variable dependiente: COLESTOT

¿Cómo sería la aplicación del método de Bonferroni para detectar *outliers* con nivel de significación 0,05

Etapa 5.- Inferencias con el modelo

Si las hipótesis del modelo se aceptan, podemos inferir (generalizar a la población) a partir de nuestro modelo

• Inferencias sobre los parámetros del modelo

Los parámetros pueden ser interpretados como medidas de la magnitud del efecto de cada variable explicativa sobre la variable respuesta (incremento en y por unidad de incremento en x_i, ajustado por el resto). Las inferencias posibles son:

Prueba de hipótesis para comprobar la significación del efecto de cada variable:

$$H_0 : \beta_i = 0 \qquad H_a : \beta_i \neq 0 \qquad i = 1,\ldots,k$$

resuelta a través del estadístico: $t = \dfrac{\hat{\beta}_i}{\sqrt{Var(\hat{\beta}_i)}}$ (t de student n-k-1 gl)

Intervalo de confianza para cada β_i

$$I_{1-\alpha}(\beta_i) = \left[\hat{\beta}_i \pm t_{1-\alpha/2}\sqrt{Var(\hat{\beta}_i)}\right]$$

con $t_{1-\alpha/2}$ coeficiente de una t de student con n-k-1 gl

Las tablas adjuntas presentan los resultados de las inferencias sobre los parámetros del modelo (pruebas t e intervalos de confianza). Interpretelos

Coeficientes[a]

Modelo		t	Sig.
	(Constante)	4,516	,000
	EDAD	4,247	,000
	QUETELET	2,784	,006

a. Variable dependiente: COLESTOT

Coeficientes[a]

Modelo		Intervalo de confianza para B al 95%	
		Límite inferior	Límite superior
1	(Constante)	60,856	155,225
	EDAD	,530	1,449
	QUETELET	,789	4,620

a. Variable dependiente: COLESTOT

• Inferencias sobre las predicciones

Compare los intervalos de predicción. Justifique las diferencias que observe entre ellos

Para un valor concreto de x (conjunto de valores para las variables explicativas), digamos

$$\mathbf{x^0 = (x_1^0, x_2^0, \ldots, x_k^0)}$$

podemos realizar dos clases de predicciones:

Predicción individual. Se trata de pronosticar el valor de la variable respuesta para **un sujeto** con valores x^0. La solución se obtendrá a través del intervalo de confianza para la predicción individual

En el ejemplo, para un sujeto de Edad=59 y Quetelet=30, tenemos

$$\mathbf{I_{0,95}(y) = [156,1 ; 339,2]}$$

Predicción a la media . Se trata de pronosticar elvalor de la media de la variable respuesta para **la población de sujetos** con valores x^0. La solución se obtendrá a través del intervalo de confianza para la predicción a la media

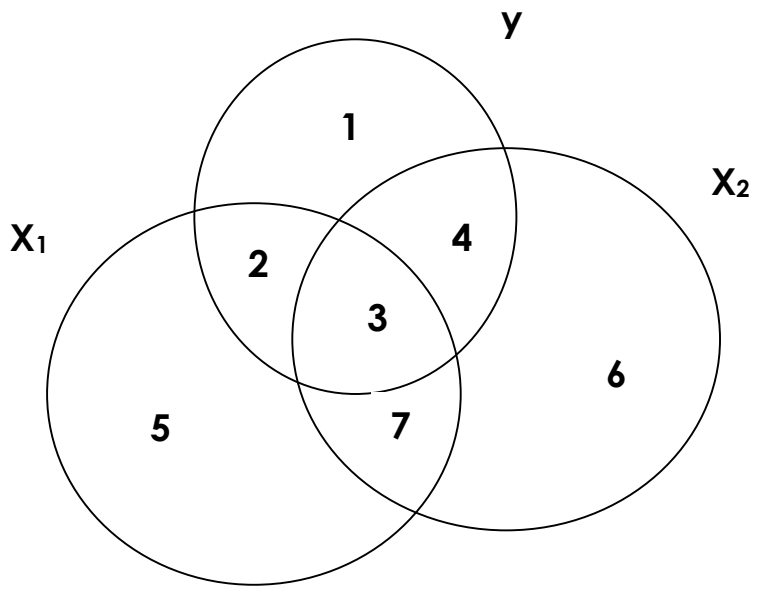
En el ejemplo, para un sujeto de Edad=59 y Quetelet=30, tenemos

$$\mathbf{I_{0,95}(\mu_y) = [238,5 ; 256,7]}$$

Algunas cuestiones adicionales

La multicolinealidad

• El grado de multicolinealidad define el grado en el que alguna o algunas variables explicativas se interrelacionan. En la mayoría de los estudios observacionales las variables explicativas presentan un cierto grado de relación, hecho que suele formar parte de los objetivos de análisis (detección de confusiones, interacciones, etc.), no siendo esto un problema sino un objetivo de análisis. Así, en la figura adjunta podemos observar una representación de la distribución de variabilidades explicadas y no explicadas en un modelo de regresión con dos variables explicativas



Variabilidad total y = 1+2+3+4

Variabilidad explicada por x_1 = 2+3

Variabilidad explicada por x_2 = 3+4

Variabilidad explicada por x_1 y x_2 = 2+3+4

Variabilidad residual no explicada =1

¿Qué región del gráfico adjunto representa el grado de multicolinealidad entre las variables x_1 y x_2?

• Una forma de evaluar el grado de multicolinealidad será comparando medidas que reflejen la capacidad explicativa de cada variable sóla con la capacidad explicativa del modelo que añade la otra variable. A través del coeficiente R^2 tendremos, por ejemplo:

$$R^2_{y/x_1} = \text{Coeficiente de determinación del modelo sólo con } x_1$$

$$R^2_{y/x_1,x_2} = \text{Coeficiente de determinación del modelo con } x_1 \text{ y } x_2$$

pudiendo expresar la capacidad explicativa que añade x_2 a través de la diferencia en R^2:

$$H = R^2_{y/x_1,x_2} - R^2_{y/x_1}$$

y contrastando la significación del incremento de capacidad explicativa de j variables

$$H_0 : H = 0 \qquad H_a : H \neq 0$$

a través del estadístico F de cambio, para el contraste de hipótesis:

$$F = \frac{H/j}{1 - R^2/n - k - 1}$$

con R^2 determinación con todas las variables (k)

Las tablas adjuntas muestran resultados de modelos de regresión añadiendo variables. Discuta el incremento de variabilidad explicada y su significación

Modelo	R cuadrado	Estadísticos de cambio: Cambio en R cuadrado	Cambio en F	gl1	gl2	Sig. del cambio en F
Uno[a]	,146	,146	33,635	1	197	,000
Dos[b]	,178	,033	7,753	1	196	,006

a. Variables predictoras: (Constante), EDAD

b. Variables predictoras: (Constante), EDAD, QUETELET

Modelo	R cuadrado	Estadísticos de cambio: Cambio en R cuadrado	Cambio en F	gl1	gl2	Sig. del cambio en F
Uno[a]	,103	,103	22,554	1	197	,000
Dos[b]	,178	,076	18,036	1	196	,000

a. Variables predictoras: (Constante), QUETELET

b. Variables predictoras: (Constante), QUETELET, EDAD

• La multicolinealidad es un problema cuando es extrema. De hecho es imposible estimar los parámetros del modelo cuando alguna de las variables explicativas es combinación lineal de las demás. En general, la multicolinealidad elevada se traduce en:

- Estimaciones poco precisas (mucho error estándar)
- Valores p elevados en las pruebas de hipótesis

La detección del grado de colinealidad puede realizarse a través de:

$R^2_{x_j / x_1 \ldots x_k}$ = Coeficiente de determinación de la regresión de x_j sobre el resto

Factor de inflación de la varianza de la variable x_j (j=1,...,k)

$$FIV_j = \frac{1}{1 - R^2_{x_j / x_1 \ldots x_k}}$$

Tolerancia de la variable x_j (j=1,...,k)

$$Tolerancia_j = \frac{1}{FIV_j} = 1 - R^2_{x_j / x_1 \ldots x_k}$$

Se habla de elevada colinealidad cuando $R^2_{x_j / x_1 \ldots x_k} > 0{,}90$, $FIV_j > 10$, $Tolerancia_j < 0{,}10$

Evalue la colinealidad para el modelo de la tabla adjunta

Coeficientes[a]

Modelo		t	Sig.	Estadísticos de colinealidad	
				Tolerancia	FIV
1	(Constante)	4,516	,000		
	EDAD	4,247	,000	,833	1,200
	QUETELET	2,784	,006	,833	1,200

a. Variable dependiente: COLESTOT

En la tabla adjunta se muestra resultados con las variables Quetelet y Quetelet^2. Discuta la colinealidad y su efecto en los resultados

Coeficientes[a]

Modelo		t	Sig.	Estadísticos de colinealidad	
				Tolerancia	FIV
1	(Constante)	-,488	,626		
	QUETELET	1,671	,096	,007	137,377
	QUET2	-1,269	,206	,007	137,377

a. Variable dependiente: COLESTOT

• Otra forma de evaluar la colinealidad, especialmente indicada cuando las variables del modelo han sido transformadas a unidades estandarizadas, es a través de los índices de condicionamiento (IC) y proporciones de varianza (%Var)(ver ejemplo a margen). Se identifica colinealidad elevada cuando IC > 30 y %Var > 0,9 en al menos dos variables.

Discuta la colinealidad del modelo con Edad, Quetelet y Quetelet² con los resultados de la tabla adjunta

Diagnósticos de colinealidad[a]

Dimensión	Indice de condición	Proporciones de la varianza		
		QUETELET	QUET2	EDAD
1	1,000	,00	,00	,00
2	7,996	,00	,00	,95
3	10,499	,00	,01	,00
4	212,135	1,00	,99	,04

a. Variable dependiente: COLESTOT

La selección del modelo óptimo

Existen diferentes formas de abordar la selección del modelo óptimo. En general depende del objetivo del estudio. Así, tendremos:

• *Selección forzada*. Definido el modelo con todas las variables explicativas a estudio se inspecciona los contrastes individuales reduciendo el modelo en aquellas que no son significativas. Útil como procedimiento exploratorio

• *Regresión jerárquica*. El investigador define un orden de introducción de las variables (primero el factor de riesgo, segundo variables confundientes, tercero interacciones, etc.) y evalúa las significaciones a través de los estadísticos y contrastes de cambio. Responde a una secuencia de análisis y objetivos bien definidos

• *Regresión por 'stepwise' o etapas*. Modalidad de regresión jerárquica pero dejando que las variables entren o salgan por criterios exclusivamente estadísticos. Útil cuando buscamos el mejor modelo predictivo pero puede conducir a modelos incoherentes.

• *Regresión sobre todos los subconjuntos posibles*. Sería el ideal. Se trata de construir la regresión sobre todas las combinaciones de variables explicativas, seleccionando el modelo óptimo atendiendo a razones de ajuste del modelo y coherencia explicativa. Puede ser muy costosa si el número de variables explicativas es elevado.

Los coeficientes de correlación

La magnitud de la relación lineal entre la variable respuesta y las explicativas puede ser cuantificada a través de coeficientes de correlación lineal (medidas estandarizadas entre 0 y 1):

Correlación lineal simple = $r_{y.x_i}$ **=** Grado de la relación lineal entre la variable respuesta (y) y una variable explicativa cualquiera (x_i)

Correlación multiple = $R_{y.(x_1 \ldots x_k)} = \sqrt{R^2_{y/x_1 \ldots x_k}}$ **=** Grado de la relación lineal conjunta entre la variable respuesta (y) y todo el conjunto de explicativas

Correlación parcial = $R_{y.x_i / x_1 \ldots x_{i-1}, x_{i+1}, \ldots x_k}$ **=** Grado de la relación lineal entre la variable respuesta (y) y una explicativa (x_i) ajustado por el resto de variables

Coeficiente de determinación corregido = $\overline{R}^2_{y/x_1 \ldots x_k} = 1 - \frac{\hat{s}^2_e}{\hat{s}^2_y}$. Corrige la determinación en función del número de variables explicativas

Interprete los coeficientes de determinación corregido, de correlación múltiple, simple y parciales de la tablas adjuntas

Modelo	R	R cuadrado	R cuadrado corregida
1	,422[a]	,178	,170

a. Variables predictoras: EDAD, QUETELET

		Correlaciones	
Modelo		Orden cero	Parcial
1	QUETELET	,321	,195
	EDAD	,382	,290

Variables explicativas cualitativas

• Todo lo expuesto hasta ahora es válido cuando las variables explicativas son cuantitativas. La variable respuesta debe ser siempre cuantitativa.

• ¿Podemos incluir variables cualitativas entre las explicativas? La respuesta es sí, pero con precaución. Si no realizamos niguna modificación sobre las vriables podemos violar fácilmente la hipótesis de linealidad. Como ejemplo considere la variable alcohol, definida en la página 5:

ALCOHOL = Consumo de alcohol= {0, 1, 2} = {Nunca, Bajo, Moderado/Alto}

De introducir la variable tal cual en un modelo de regresión lineal, La linealidad se traduciría en que el efecto sobre la variable respuesta por pasar de 0 (nunca) a 1 (Bajo) fuera el mismo que por pasar de 1 (bajo) a 2 (moderado/alto), algo que es cuanto menos dudoso.

• Sin embargo, si nos referimos a una variable cualitativa dicotómica (dos estados o categorías), no existe ambigüedad sobre la linealidad, puesto que sólo existe un efecto o 'salto' posible. Por ejemplo, si nos referimos a la variable:

SEXO = {1, 2} = {Hombre, Mujer}

Podemos introducir la variable tal cual, puesto que sólo hay un paso posible, de 1 (hombre) a 2 (mujer) y el parámetro de la variable representará el efecto sobre la variable respuesta por comparar mujeres vs. hombres

Compruebe y discuta que la interpretación de los parámetros de una variable cualitativa dicotómica es dependiente de la codificación elegida

• Una solución posible sería analizar los datos por separado (en cada categoría de la variabe cualitativa) pero esto es generalmente poco eficiente

Ponga algunos ejemplos de variables cualitativas dicotómicas

Procedimiento general de introducción de variables cualitativas

• Supongamos una situación con x_1=variable cualitativa, x_2=variable cuantitativa. Debemos distinguir dos casos:

Caso de variables dicotómicas

i. Codificación. Lo más sencillo es codificar la variable, x_1, con códigos que disten 1 unidad, por ejemplo:

$$x_1 = \{0,1\}$$

ii. Modelo resultante. Interpretación de parámetros.- El modelo resultante será:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

y el parámetro β_1 representa el cambio en y por comparar las dos categorías de x_1

Caso de variables con más de dos categorías

Suponga la variable 'Nivel de estudios' con 5 categorías

Nivel de estudios={Analfabeto, Leer y escribir, Primarios, Secundarios, Universitarios}

Cree estructura de dummys tomando como referencia la categoría Universitarios

• Creación de variables *dummy* o indicador. Se trata de generar tantas variables auxiliares como el número de categorías de la variable menos 1: Supongamos

x_1=Consumo de alcohol={Nunca, Bajo, Moderado/Alto}

i. Se elige una categoría de referencia, p. ej. Nunca. La elección de la categoría de referencia no influye en el ajuste y significaciones del modelo, pero si influye en la interpretación, siendo recomendable elegir la categoría de 'menor riesgo' o 'más favorable' cuando la variable respuesta representa un resultado negativo o lo contrario cuando representa un resultado positivo

Categoría de referencia = Moderado/Alto

ii. Codificación de las variables dummy. En este caso habrán 2 variables dummy (nº de categorías -1=3 – 1 = 2):

Tabla 2.- Codificación de dummys

Codificación		
CONSUMO DE ALCOHOL X_1	DUMMY X_1^1	DUMMY X_1^2
Nunca	0	0
Bajo	1	0
Moderado/Alto	0	1

iii. Modelo resultante. El modelo resultante será:

$$y = \beta_0 + \beta_1^1 x_1^1 + \beta_1^2 x_1^2 + \beta_2 x_2$$

resultando la siguiente interpretación:

β_1^1 = Incremento sobre y por comparar consumo Bajo vs. Nunca

β_1^2 = Incremento sobre y por comparar consumo Moderado/alto vs. Nunca

iv. Con variables dummy la significación de parámetros y validez del modelo se comprueba de forma habitual. Es **muy importante** tener en cuenta que el efecto de una variable debe ser evaluado con todas sus dummys. Con otras palabras, las dummys son inseparables en el modelo, no podemos introducir sólo una parte ni eliminar del modelo ninguna cuando la variable está en él.

La evaluación del conjunto de dummys puede hacerse a través de los estadísticos y significación del cambio

Compruebe la interpretación de parámetros en las variables dummy. En el ejemplo seguido, ¿cómo se puede estimar el efecto por comparar consumo de alcohol Moderado/alto vs. Bajo?

La tabla adjunta muestra los resultados al ajustar el modelo con Edad y Quetelet y el modelo que añade las dummys Alcohol1 y Alcohol2, como definidas en tabla 2. Compruebe el efecto significativo de la variable Alcohol completa

Resumen del modelo

Modelo	Estadísticos de cambio		
	Cambio en R cuadrado	Cambio en F	Sig. del cambio en F
Uno[a]	,178	21,270	,000
Dos[b]	,065	8,351	,000

a. QUETELET, EDAD

b. QUETELET, EDAD, ALCOHOL2, ALCOHOL1

El MODELO DE REGRESION LOGISTICA

Introducción. Medidas de frecuencia y asociación en análisis epidemiológico

• Suponga una cohorte de 10 individuos (tabla 3), seguidos durante diferentes tiempos (variable TIEMPO, en meses), a la espera de observar si se les presenta (incide) cierto resultado (variable EVENTO, 1 'se presenta el resultado', 0 'no se presenta el resultado), teniendo en cuenta que al inicio del seguimiento 5 de estos individuos presentan cierta característica (variable GRUPO=1) y 5 verifican otro estado en esta característica (variable GRUPO=2):

Tabla 3. Resultados en una cohorte de 10 individuos

Individuo	Tiempo	Evento	Grupo
1	10	0	1
2	5	0	1
3	5	1	1
4	4	0	1
5	2	1	1
6	10	0	2
7	5	1	2
8	5	1	2
9	2	1	2
10	2	1	2

Para revisar diferentes medidas epidemiológicas, conteste las preguntas adjuntas

¿Qué medidas utilizamos para cuantificar

a) La magnitud de ocurrencia del evento en la cohorte al final del seguimiento y sus diferencias (o asociación) con la variable GRUPO, sin tener en cuenta el tiempo específico de seguimiento de cada individuo. Calcúlelas

b) Idem a a) pero teniendo en cuenta los tiempos de seguimiento de los sujetos. Calcúlelas

c) La magnitud de ocurrencia del evento en diferentes momentos del tiempo y su asociación con el GRUPO. Calcúlelas

El modelo logístico dicotómico. Análisis simple

• Suponga una situación como la descrita en el apartado a) de la columna de aplicaciones de la página anterior:

- Seguimiento de una cohorte de n individuos durante un periodo $[t_0, t]$
- Se observa la ocurrencia (E=1) o no (E=0) de un evento E
- En el instante t_0 cada individuo puede estar expuesto (x=1) o no (x=0) a un factor de riesgo, digamos x
- Queremos estimar la probabilidad de ocurrencia del evento (p(E=1)) y su asociación con el factor x

Podemos considerar la siguiente tabla de probabilidades condicionales:

Tabla 4.- Probabilidades condicionales en función del factor x

	x=0	x=1
E=0	$p(E = 0 / x = 0)$	$p(E = 0 / x = 1)$
E=1	$p(E = 1 / x = 0)$	$p(E = 1 / x = 1)$

Revise el concepto de *odds*. Aplíquelo al evento E condicionado por x (odds(E/x=1); odds(E/x=0)

Revise el concepto de *odds ratio*. Aplíquelo al evento E condicionado por x

¿Recuerda el concepto de *cross product ratio* (cociente de productos cruzados)?

• El modelo logístico consiste en modelizar las probabilidades del evento como:

¿Cómo serían las expresiones de las probabilidades condicionales de E dado x si x estuviera codificada con códigos 1 y 2?

(1) $$p(E = 1/x) = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}} = \frac{e^{(\beta_0+\beta_1 x)}}{1+e^{(\beta_0+\beta_1 x)}} \qquad (x=0,1)$$

o, equivalentemente

$$logit(p(E = 1/x)) = log\,odds(E = 1/x) =$$

(2)

$$\log\left[\frac{p(E = 1/x)}{1 - p(E = 1/x)}\right] = \beta_0 + \beta_1 x$$

siendo log el logaritmo neperiano (de base e). Utilizando la formulación de (1), podemos expresar las probabilidades condicionales de la tabla 4 como:

$$p(E = 1/x = 1) = \frac{e^{(\beta_0+\beta_1)}}{1+e^{(\beta_0+\beta_1)}} \qquad p(E = 0/x = 1) = \frac{1}{1+e^{(\beta_0+\beta_1)}}$$

$$p(E = 1/x = 0) = \frac{e^{(\beta_0)}}{1+e^{(\beta_0)}} \qquad p(E = 0/x = 0) = \frac{1}{1+e^{(\beta_0)}}$$

Pudiendo comprobar que el ***odds ratio*** de E condicionado por x es:

Obtenga la interpretación del parámetro β_0 del modelo de regresión logística

$$OR_{E/x} = \frac{p(E=1/x=1)\,/\,p(E=0/x=1)}{p(E=1/x=0)\,/\,p(E=0/x=0)} =$$

$$= \frac{\dfrac{e^{(\beta_0+\beta_1)}}{1+e^{(\beta_0+\beta_1)}} \Big/ \dfrac{1}{1+e^{(\beta_0+\beta_1)}}}{\dfrac{e^{(\beta_0)}}{1+e^{(\beta_0)}} \Big/ \dfrac{1}{1+e^{(\beta_0)}}} = \frac{e^{(\beta_0+\beta_1)}}{e^{(\beta_0)}} = e^{\beta_1}$$

Obteniendo una relación funcional directa entre odds ratio y coeficiente del modelo:

$$\boxed{OR_{E/x} = e^{\beta_1}} \qquad (3)$$

El modelo logístico. Generalización a k variables explicativas

En el Anexo 2 puede consultar los datos de las variables descritas en la tabla 5.

• La generalización del modelo para un vector de k variables explicativas, digamos $x=(x_1, x_2, ..., x_k)$ tiene la formulación:

(4) $$p(E = 1/x) = \frac{1}{1+e^{-(\beta_0+\beta_1x_1+...+\beta_kx_k)}} = \frac{e^{(\beta_0+\beta_1x_1+...+\beta_kx_k)}}{1+e^{(\beta_0+\beta_1x_1+...+\beta_kx_k)}}$$

o, equivalentemente

A partir de la variable respuesta INFARTO, formule un modelo de regresión logística como función de las variables TABACO, ECG y PAS

$$logit(p(E = 1/x)) = logodds(E = 1/x) =$$

(5)

$$log\left[\frac{p(E = 1/x)}{1-p(E = 1/x)}\right] = \beta_0 + \beta_1x_1 + ... + \beta_kx_k$$

Siendo $x_1, ..., x_k$ variables categóricas o cuantitativas. A modo de ejemplo considere observaciones sobre 99 sujetos de las variables de la tabla adjunta:

Tabla 5.- Variables en un estudio de factores de riesgo en infarto

INFARTO	= 0 'No' 1 'Si'	EDAD= En años	
TABACO	= 0 'No fuma' 1 'Fumador'	SEXO = 1 'Hombre' 2 'Mujer'	
PAS	= Presión sistólica (mmHg)	PAS1 = 1 '<130' 2 '130-150' 3 '>150'	
ECG	= 0 'Normal' 1 'Anormal'	PAS2 = 1 '<150' 2 '>=150'	

• La generalización en la interpretación de parámetros respecto al modelo simple puede realizarse considerando la siguiente situación: Sean

$\mathbf{x} = (x_1,\ldots,x_i,\ldots,x_k)$ vector de valores o categorías de las variables explicativas (un perfil determinado, por ejemplo no fumador, ecg normal, pas=120 mmHg)

$\mathbf{x}^* = (x_1^*,\ldots,x_i^*,\ldots,x_k^*)$ vector de otros valores o categorías de las variables explicativas (otro perfil determinado, por ejemplo fumador, con ecg anormal y pas=160 mmHg)

Se define el *odds ratio* de asociación entre la variable respuesta dicotómica (infarto) y la variable dicotómica (x, x*) como el odds ratio de la tabulación:

$$OR_{E/(x,x^*)} = \frac{p(E=1/x^*)/p(E=0/x^*)}{p(E=1/x)/p(E=0/x)}$$

	x	x*
E=0	$p(E=0/x)$	$p(E=0/x^*)$
E=1	$p(E=1/x)$	$p(E=1/x^*)$

Medida que representa el grado de asociación entre la respuesta (infarto) y dos perfiles distintos de las explicativas ([no fumador, ecg normal, pas=120] vs. [fumador, ecg anormal, pas=160])

¿Cuántos perfiles distintos existen con las variables explicativas TABACO, ECG y SEXO? Exprese algunos de ellos. ¿Y cuántos OR de asociación entre dos perfiles se pueden construir?

Si alguna de las variables explicativas es continua, por ejemplo PAS, ¿Cuántos perfiles distintos existen?

• Si modelizamos las probabilidades de E según el modelo logístico (4) tendremos:

$$OR_{E/(x,x^*)} = \frac{p(E=1/x^*)\,/\,p(E=0/x^*)}{p(E=1/x)\,/\,p(E=0/x)} =$$

$$= \frac{\dfrac{e^{\beta_0+\beta_i x_1^*+\ldots+\beta_k x_k^*}}{1+e^{\beta_0+\beta_i x_1^*+\ldots+\beta_k x_k^*}}}{\dfrac{1}{1+e^{\beta_0+\beta_i x_1^*+\ldots+\beta_k x_k^*}}} \Big/ \frac{\dfrac{e^{\beta_0+\beta_i x_1+\ldots+\beta_k x_k}}{1+e^{\beta_0+\beta_i x_1+\ldots+\beta_k x_k}}}{\dfrac{1}{1+e^{\beta_0+\beta_i x_1+\ldots+\beta_k x_k}}} = \frac{e^{\beta_0+\beta_1 x_1^*+\ldots+\beta_k x_k^*}}{e^{\beta_0+\beta_1 x_1+\ldots+\beta_k x_k}} = e^{\sum_{i=1}^{k}\beta_i(x_i^*-x_i)}$$

$$\boxed{OR_{E/(x,x^*)} = e^{\sum_{i=1}^{k}\beta_i(x_i^*-x_i)} = \prod_{i=1}^{k} e^{\beta_i(x_i^*-x_i)}} \qquad (6)$$

Pudiendo expresar el odds ratio como función de los parámetros del modelo a través de los cambios en las variables explicativas. Nótese que no depende de β_0.

Escriba la expresión del odds ratio de asociación entre infarto y los perfiles (no fuma, ecg normal, pas=120) vs. (fuma, ecg anormal, pas=150)

Escriba la expresión del odds ratio de asociación entre infarto y los perfiles en los que sólo cambia la variable tabaco (no fuma, ecg, pas) vs. (fuma, mismo ecg, mismo pas). ¿Cómo se llama ese odds ratio? Interprételo

El modelo de regresión logística es considerado un modelo de efectos multiplicativos. Discuta este concepto a partir de la expresión (6)

Requerimientos en el modelo de regresión logística

• A diferencia del modelo de regresión lineal, las inferencias no requerirán suposición distribucional alguna. Se dirá que las inferencias son asintóticas, es decir, válidas para un n suficientemente grande. De acuerdo con la ecuación (6) se supone composición multiplicativa de efectos (aditiva en la escala logarítmica)

En su opinión, ¿cuál o cuales de las variables restantes son candidatas a influir sobre la probabilidad de infarto? ¿Se le ocurren otras que no estén entre las estudiadas?

Construcción de un modelo de regresión logística. Etapas

• Suponga el ejemplo descrito anteriormente, con observaciones sobre 99 individuos de las variables descritas en la tabla 5

Etapa 1: Especificación de variables y modelo propuesto

Se desea averiguar si las variables TABACO y PAS influyen sobre la probabilidad de INFARTO. Si denotamos por

E = 1 'Infarto' 0 'No infarto'

x_1 = 0 'No fuma' 1 'Fumador'

x_2 = PAS (mmHg)

$p(x)=p(E=1/x_1,x_2)$

Modelo propuesto

$$\log\frac{p(x)}{1-p(x)} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Etapa 2: Estimación del modelo

¿Recuerda en qué se traducía gráficamente el criterio de mínimos cuadrados en regresión lineal simple?

A partir de los datos disponibles, una muestra aleatoria de n observaciones de las variables:

$$\{E_i, x_{1i}, x_{2i}, \cdots, x_{ki}\}_{i=1}^{n}$$

Se trata de calcular los estimadores muestrales de los parámetros del modelo:

$(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k)$ estimadores de los parámetros $(\beta_0, \beta_1, \beta_2, \ldots, \beta_k)$

El método de estimación utilizado es el de máxima verosimilitud. Este método presenta diferencias con el método de mínimos cuadrados.

El principio de máxima verosimilitud

• Partiendo de la idea de que a partir de los datos muestrales queremos obtener una función que nos produzca un valor (estimador muestral) que se aproxime al de un parámetro poblacional, de forma que sea 'aceptable' (coherente y de buenas propiedades) y 'operativo' (permita su cálculo matemático), podemos utilizar diferentes criterios. El criterio de mínimos cuadrados utilizado en regresión lineal se basa en argumentos geométricos, partiendo de la idea de minimizar:

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Notas

• Frente a esta idea geométrica, el criterio de máxima verosimilitud se basa en la distribución de probabilidad de los datos observados (dependiente del parámetro), construyendo como estimador del parámetro aquel valor que asigna mayor probabilidad a los datos observados

• Para comprender este procedimiento, supongamos la siguiente situación:

'Se desea averiguar el valor poblacional de la prevalencia de un problema de salud, digamos θ. Para ello extremos una muestra de tamaño n=5 y en cada individuo se observa si posee o no el problema de salud, resultando que 4 de ellos se ven afectados. Para simplificar suponamos que θ sólo puede valer 0,2 o 0,6 (en realidad será cualquier valor entre 0 y 1)'

El criterio de máxima verosimilitud requiere lo siguiente:

i. Suponer una distribución de probabilidad para los datos, función de el (los) parámetro(s) de interés. Si denotamos por X=nº de personas que poseen el problema entre las 5 observadas, tenemos que:

$$X \approx \text{Binomial}(n, \theta) = \text{Bi}(n = 5, \theta)$$

$$p(X = k) = \binom{n}{k} \theta^k (1-\theta)^{n-k} = \binom{5}{k} \theta^k (1-\theta)^{5-k}$$

con θ=0,2 o 0,6. Esta función es la verosimilitud, una vez observado k

ii. Decidir como estimación de θ aquel valor que maximiza la la función anterior. En la tabla adjunta se observa el valor de la función de verosimilitud para los diferentes valores de k. ¿Cuál es el valor de θ que hace más verosímil (maximiza la función de verosimilitud) al resultado k=4?

¿Recuerda cuál es el estimador máximo verosímil de la media de una variable?¿En qué modelo de probabilidad de basa?

Tabla 6.- Valores de la función de verosimilitud para diferentes resultados (k=0,...,6) y θ=0,2 o 0,6

K / θ	p(X=k)					
	k=0	**k=1**	**k=2**	**k=3**	**k=4**	**k=5**
θ=0,2	0,328	0,409	0,205	0,051	0,007	<0,001
θ=0,6	0,010	0,077	0,230	0,346	0,259	0,078

En general, para $0<\theta<1$, se trata de maximizar la función de verosimilitud:

$$l(x,\theta) = \binom{n}{k}\theta^k(1-\theta)^{n-k}$$ o equivalentemente

$$\log l(x,\theta) = k\log\theta + (n-k)\log(1-\theta)$$, derivando e igualando a 0 se tiene:

$$\frac{\delta}{\delta\theta}\log l(x,\theta) = \frac{k}{\theta} - \frac{n-k}{1-\theta} = 0 \Rightarrow \theta = \frac{k}{n} \Rightarrow \boxed{\hat{\theta} = \frac{4}{5} = 0{,}80}$$

Notas

Funciones de verosimilitud en el modelo logístico

Existen dos funciones de verosimilitud posibles en la situación de aplicación del modelo logístico:

• *Verosimilitud no condicional*

Se trata de expresar la probabilidad del conjunto de datos observados sin ninguna restricción. Así, si partimos de una muestra de n sujetos, cada uno de ellos con un vector de variables explicativas $x_l = (x_{l1}, x_{l2},...,x_{lk})$, de los que

n_1 casos (E=1), asociados con los vectores de variables explicativas $x_1,...,x_{n1}$
n_0 no casos (E=0), asociados con el resto de vectores $x_{n1+1},...,x_n$

La verosimilitud es la probabilidad de obtener n_1 casos asociados con los n_1 vectores y n_0 no casos asociados con el resto de vectores, que, de acuerdo con un modelo binomial será:

$$l(x,\beta) = \prod_{l=1}^{n_1} p(E=1/x_l) \prod_{l=n_1+1}^{n} p(E=0/x_l) = \frac{\prod_{l=1}^{n_1} e^{\beta_0+\beta_1 x_{l1}+...+\beta_k x_{lk}}}{\prod_{l=1}^{n}(1+e^{\beta_0+\beta_1 x_{l1}+...+\beta_k x_{lk}})}$$

Notas

• *Verosimilitud condicional*

Ordenados los n sujetos y sus correpondientes vectores de variables explicativas:

Casos			No casos		
1	...	n_1	n_1+1	...	n
$x_{1,1}$		$x_{n1,1}$	$X_{n1+1,1}$		$x_{n,1}$
$x_{1,2}$		$x_{n1,2}$	$x_{n1+1,2}$		$x_{n,2}$
.		.	.		.
.		.	.		.
$x_{1,k}$		$x_{n1,k}$	$x_{n1+1,k}$		$x_{n,k}$

Sabiendo que n_1 han sido casos y n_0 no casos, la verosmilitud representa la probabilidad de que las n_1 primeras columnas correspondan a los casos. Se la llama condicional por ser condicional a que el numero de casos ha sido n_1 y su forma es:

$$l(x,\beta) = \frac{\prod_{l=1}^{n_1} e^{\beta_1 x_{l1}+\ldots+\beta_k x_{lk}}}{\sum_{u}\left(\prod_{l=1}^{n}(1+e^{\beta_1 x_{l1}+\ldots+\beta_k x_{lk}})\right)}$$

donde u representa un índice que recorre todas las combinaciones posibles de n elementos tomados de n_1 en n_1. Nótese que el parámetro β_0 no está en la función

• *Usos de ambas verosimilitudes*

- La verosimilitud no condicional es la forma más natural de expresión de la probabilidad de los datos. El número de casos y no casos no está prefijado

- La verosimilitud condicional suele reservarse a situaciones en las que la no condicional produce resultados poco eficientes

- Esto sucede sobre todo en estudios con diseño apareado, en los que al analizar la información de cada estrato, generalmente definido por el conjunto caso-control(es), la verosimilitud no condicional conduce a estimaciones menos precisas

- Cuando el número de casos y no casos es alto, la verosimilitud condicional puede resultar impracticable debido a que su denominador trabaja con combinaciones de expresiones

- Resumiendo, la condicional suele utilizarse en estudios retrospectivos con apareamiento

El proceso de estimación

Recuerde los conceptos de varianza y error estándar de un estimador y el de covarianza entre estimadores

• Una vez hemos decidido la función de verosimilitud a utilizar, el criterio de máxima verosimilitud produce los siguientes resultados:

Tabla 7.- Resultados en la estimación máximo-verosímil

$\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k)$	Estimadores de los parámetros
$l(\hat{\beta})$	Valor de la función de verosimilitud para el modelo
$\log l(\hat{\beta})$	Log-verosimilitud. Valor del logaritmo de la verosimilitud
$V = \begin{pmatrix} s^2_{\hat{\beta}_0} & & \\ \cdot & \cdot & \\ \cdot & & \cdot \\ s_{\hat{\beta}_0\hat{\beta}_k} & \cdot \; \cdot & s^2_{\hat{\beta}_k} \end{pmatrix}$	Matriz simétrica de varianzas covarianzas de los estimadores de los parámetros del modelo
$s^2_{\hat{\beta}_i} = Var(\hat{\beta}_i)$	Elemento i-ésimo de la diagonal de la matriz V
$s_{\hat{\beta}_i} = Error\ Estándar(\hat{\beta}_i)$	
$s_{\hat{\beta}_i\hat{\beta}_j} = Cov(\hat{\beta}_i, \hat{\beta}_j)$	Covarianza entre los estimadores de los parámetros

• Como consecuencia del proceso de estimación dispondremos del modelo logístico estimado. Así, para el modelo propuesto en la etapa 1, con los datos del anexo 2, estimamos el siguiente modelo (verosimilitud no condicional):

Interprete los parámetros del modelo. Tradúzcalos a *odds ratios* de asociación entre infarto y tabaco y pas

$$\log \frac{p(x)}{1-p(x)} = -8{,}950 + 2{,}246x_1 + 0{,}054x_2$$

Etapa 3: Bondad de ajuste del modelo

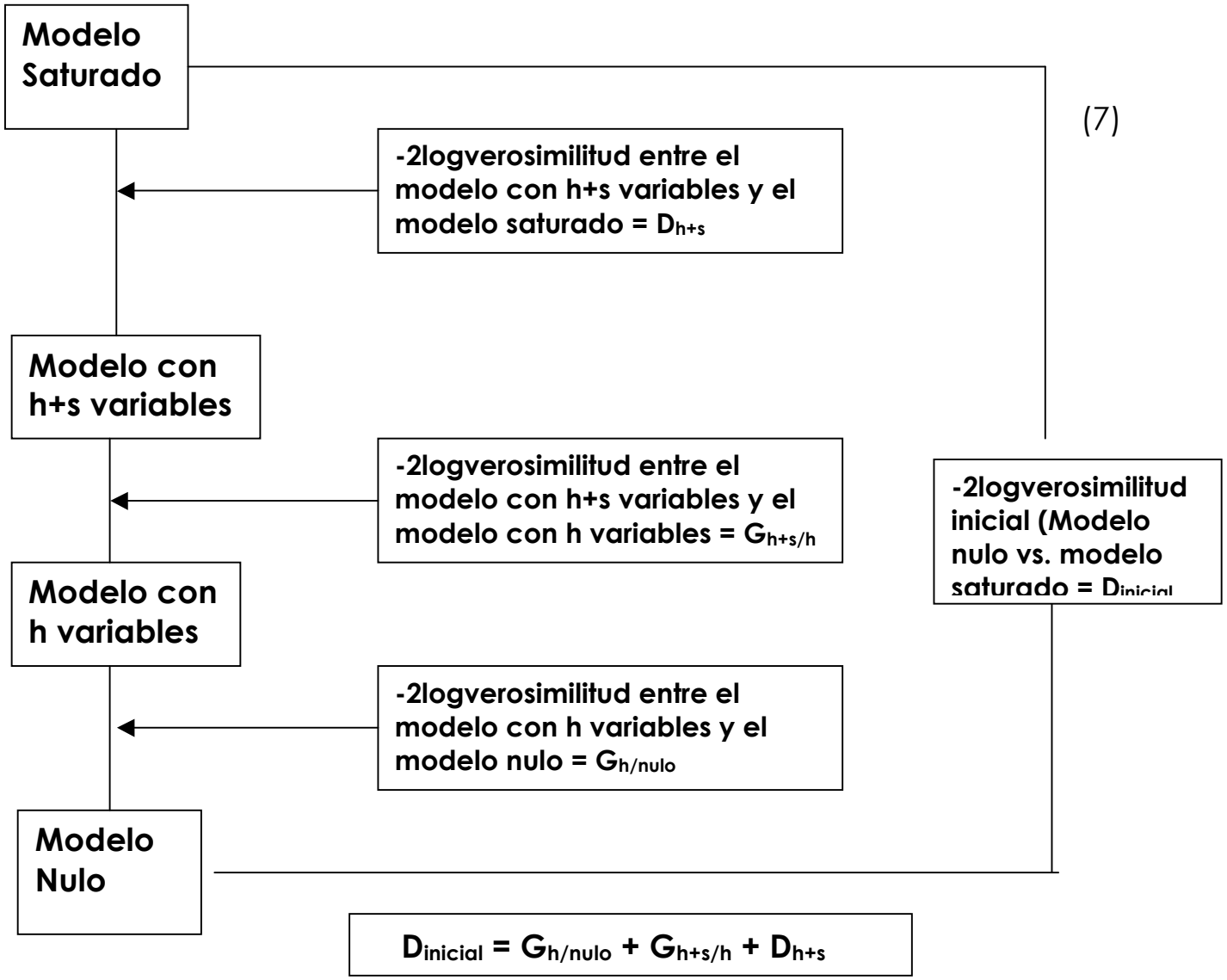
• La bondad de ajuste del modelo se evaluará a través de la log-verosimilitud, concretamente a través de la función:

$$-2\log l(\hat{\beta}) = -2\log \text{verosimilitud}$$

debiendo definir el concepto de:

Modelo saturado: Modelo que reproduce exactamente los datos. Tiene tantos parámetros como observaciones. Este modelo tendrá un valor de $-2\log l(\hat{\beta})$ que será el mínimo posible, 0. Cualquier otro modelo con menos parámetros tendrá mayor valor de –2logverosimilitud

La bondad de ajuste de modelos se hará de acuerdo con el siguiente esquema (el termómetro de bondad):



Con los datos del ejemplo propuesto en la etapa 1, se tiene lo siguiente:

-2logVerosimilitud inicial 137,243

-2log verosimilitud modelo tabaco y pas vs. modelo saturado 87,007

Sitúe estos valores en un esquema como (7)
¿Cuánto vale la distancia -2logverosimilitud entre el modelo tabaco y pas y el modelo nulo?

• El esquema (7) se basa en los siguientes elementos de evaluación de la bondad de un modelo:

$$D_s = -2\log\left[\frac{\text{verosimilitud modelo con s parámetros}}{\text{verosimilitud modelo saturado}}\right] \quad (8)$$

Esta cantidad es conocida generalmente como *deviance* (discrepancia) del modelo a estudio, y representa la distancia en bondad de ajuste que le falta al modelo a estudio para llegar a un modelo saturado que reproduce perfectamente los datos

$$G_{h+s/h} = -2\log\left[\frac{\text{veros. modelo con h + s parámetros}}{\text{veros. modelo con h parámetros}}\right] = D_s - D_{h+s} \quad (9)$$

Esta cantidad es una medida de la ganancia en capacidad explicativa por añadir s parámetros (variables) a un modelo con h parámetros (variables)

El contraste de hipótesis:

H_0: Las s variables nuevas no incrementan significativamente el ajuste del modelo
H_a: Las s variables nuevas incrementan significativamente el ajuste del modelo

Puede ser resuelto a través del estadístico $G_{h+s/h}$, que se distribuye bajo H_0 como una Ji-cuadrado con s grados de libertad

Establezca cómo evaluar la significación de la bondad de un modelo cualquiera en base a (8) y (9). Aplíquelo al modelo de tabaco y pas del ejemplo.

Etapa 4.- Inferencias con el modelo

• A partir de los elementos que nos produce la estimación máximo-verosímil (ver tabla 7) podemos realizar las siguientes inferencias:

• Inferencias sobre los parámetros del modelo

Los parámetros pueden ser interpretados como medidas de la magnitud del efecto de cada variable explicativa sobre la variable respuesta. Las inferencias posibles son:

Prueba de hipótesis para comprobar la significación de cada variable:

$$H_0 : \beta_i = 0 \qquad H_a : \beta_i \neq 0 \qquad i = 1,\ldots,k$$

resuelta a través del estadístico de Wald:

$$z = \frac{\hat{\beta}_i}{s_{\hat{\beta}_i}} \quad \text{(normal bajo } H_0\text{)} \quad \text{o} \qquad \chi^2 = \left(\frac{\hat{\beta}_i}{s_{\hat{\beta}_i}}\right)^2 \quad \text{(Ji-cuadrado con 1 gl)}$$

Intervalo de confianza para cada β_i

$$I_{1-\alpha}(\beta_i) = \left[\hat{\beta}_i \pm z_{1-\alpha/2} s_{\hat{\beta}_i}\right]$$

con $z_{1-\alpha/2}$ coeficiente de una normal

La tabla adjunta muestra los resultados obtenidos en el proceso de estimación del modelo con variables tabaco y pas. Discuta la significación de las variables y construya los intervalos de confianza de nivel 95% para los parámetros del modelo (ET= Error estándar de la estimación)

Variables en la ecuación

		B	E.T.	Wald
a	TABACO	2,246	,568	15,648
	PAS	,054	,017	10,305
	Constante	-8,950	2,375	14,197

a. Variable(s) : TABACO, PAS.

• Intervalo de confianza para una combinación lineal de parámetros del modelo

Dada una combinación lineal de los parámetros del modelo, con $\{a_i\}_{i=1}^{k}$ constantes conocidas:

$$C = \sum_{i=1}^{k} a_i \beta_i$$

El intervalo de confianza para C puede ser obtenido a través de la expresión:

$$I_{1-\alpha}(C) = \left[\hat{C} \pm z_{1-\alpha/2}\sqrt{Var(\hat{C})}\right] \qquad (10)$$

con:

$\hat{C} = \sum_{i=1}^{k} a_i \hat{\beta}_i$ estimador de C, combinación lineal de estimadores de los parámetros del modelo

$Var(\hat{C}) = \sum_{i=1}^{k} a_i^2 Var(\hat{\beta}_i) + 2\sum_{i<j}\sum_{j\neq i} a_i a_j Cov(\hat{\beta}_i, \hat{\beta}_j)$ (11) varianza estimada

con varianzas y covarianzas obtenidas de la matriz de varianzas-covarianzas (tabla 7)

Sabiendo que la correlación entre dos estimadores de parámetros es su covarianza dividida por el producto de sus errores estándar, a partir de la matriz de correlaciones adjunta, y los errores estándar utilizados en la pag. anterior, calcule la covarianza entre los estimadores de los parámetros de las variables tabaco y pas

	TABACO	PAS
TABACO	1,000	-,623
PAS	-,623	1,000

Desarrolle la expresión para el cálculo del intervalo de confianza de la predicción del log-odds de infarto para fumadores con pas de 160 mmHg.

• Intervalo de confianza para odds ratios

De acuerdo con la expresión general (6) para un odds ratio de asociación de la variable respuesta con dos perfiles de las explicativas:

$$OR_{E/(x,x^*)} = e^{\sum_{i=1}^{k} \beta_i (x_i^* - x_i)} = e^{C}$$

El intervalo de confianza para el odds ratio puede ser obtenido como:

$$I_{1-\alpha}(OR_{E/(x,x^*)}) = I_{1-\alpha}(e^{C}) = \left[e^{\hat{C} - z_{1-\alpha/2}\sqrt{Var(\hat{C})}} ; e^{\hat{C} + z_{1-\alpha/2}\sqrt{Var(\hat{C})}} \right]$$

siendo $\hat{C} = \sum_{i=1}^{k} \hat{\beta}_i (x_i^* - x_i)$ combinación lineal de estimadores de los parámetros del modelo, y $\mathbf{Var(\hat{C})}$ obtenida como en (11)

Utilizando la información de la tabla adjunta, calcule el intervalo de confianza de nivel 95% para el odds ratio entre infarto y tabaco ajustado por pas (pas constante)

Variables en la ecuación

		B	E.T.	Wald
a	TABACO	2,246	,568	15,648
	PAS	,054	,017	10,305
	Constante	-8,950	2,375	14,197

a. Variable(s) : TABACO, PAS.

Calcule el odds ratio y su intervalo de confianza al 95% entre tabaco y los perfiles (no fumador, pas=120) frente a (fumador, pas=160). Utilice la covarianza entre tabaco y pas calculada en la página anterior

Algunas cuestiones adicionales

Recuerde, ¿cuál es la condición necesaria y suficiente para que una variable sea confundiente del efecto entre otras dos?

Confusión e interacción. Detección con regresión logística

• Suponga 3 variables cualitativas dicotómicas:

E = 0 'No aparición evento' 1 'Se produce evento'
F = 0 'No expuesto a factor' 1 'Expuesto a factor'
C= 0 'Factor externo ausente' 1 'Factor externo presente'

Cuya distribución de frecuencias conjunta al observar 200 sujetos, considerando el efecto de C es:

C=1			
	F=1	**F=0**	
E=1	17	53	70
E=0	3	27	30
	20	80	100

$OR_{E/F}$ =2,89

C=0			
	F=1	**F=0**	
E=1	23	7	30
E=0	37	33	70
	60	40	100

$OR_{E/F}$=2,93

Mientras que si ignoramos el efecto de C es:

	F=1	**F=0**	
E=1	40	60	100
E=0	40	60	100
	80	120	200

$OR_{E/F}$=1,00

La variable C es confundiente del efecto de F sobre E. Si la tenemos en cuenta existe asociación y es la misma para cualquier estado de C. Si no la tenemos en cuenta no detectamos asociación

• Suponga ahora que la distribución conjunta ignorando el efecto de C es:

	F=1	F=0	
E=1	40	60	100
E=0	20	80	100
	60	140	200

$OR_{E/F}$=2,66

Mientras que al considerar el efecto de C, tenemos:

C=0			
	F=1	**F=0**	
E=1	4	16	20
E=0	4	16	20
	8	32	40

$OR_{E/F}$ =1,00

C=1			
	F=1	**F=0**	
E=1	36	44	80
E=0	16	64	80
	52	108	160

$OR_{E/F}$=3,27

Comprobando que sólo existe asociación entre E y F cuando C=1.

La variable C interacciona con E y F. El efecto de interacción consiste en la modificación de la asociación entre E y F según el estado de C. En este caso sólo existe asociación entre E y F cuando C=1

Si hay interacción ¿puede haber confusión?. Jerarquice los efectos

• La detección de la confusión a través de la regresión logística puede hacerse valorando los cambios en los coeficientes de las variables o su transformación en odds ratios de no contemplar en el modelo a incluuir en el modelo la(s) variable(s) confundientes. Generalmente la valoración se basa en una combinación del porcentaje de cambio y la inspección de los intervalos de confianza (Vea ejemplo adjunto)

• Sin embargo, la detección de la interacción requiere modificar la estructura del modelo, introduciendo términos que permitan su detección y cuantificación. Así, para una situación con dos variables explicativas, tendremos:

Modelo sin interacción: $$\log\frac{p(x)}{1-p(x)} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Modelo con interacción: $$\log\frac{p(x)}{1-p(x)} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12}(x_1 \cdot x_2)$$

El parámetro β_{12} captura la existencia de interacción, representada en la variable producto ($x_1.x_2$). Si este parámetro es significativo, habremos detectado una interacción significativa, debiendo incorporar ésta a las estimaciones de los odds ratios (Vea ejemplo adjunto)

Ejemplo de valoración de confusión. Considere el modelo ya estudiado con efectos de tabaco y pas sobre infarto. Se desea saber si el efecto del tabaco está confundido por la pas. Discútalo a partir de las tablas adjuntas:

Variables en la ecuación

	B	Sig.	Exp (B)	I.C. 95,0% para EXP(B)	
				Inferior	Superior
[a] TABACO	2,540	,000	12,686	4,554	35,336
Constante	-1,575	,000	,207		

a. Variable(s) TABACO.

Variables en la ecuación

	B	Sig.	Exp (B)	I.C. 95,0% para EXP(B)	
				Inferior	Superior
[a] TABACO	2,246	,000	9,447	3,105	28,743
PAS	,054	,001	1,056	1,021	1,091
Constante	-8,950	,000	,000		

a. Variable(s) TABACO, PAS.

Para las variables del modelo anterior, infarto, tabaco y pas, construya el modelo con interacción (entre tabaco, pas e infarto) y deduzca las expresiones del odds ratio de asociación entre infarto y tabaco modificado por pas. Idem para el odds ratio entre infarto y pas modificado por tabaco.

La selección del modelo óptimo

Notas

Las estrategias no son diferentes a las contempladas en regresión lineal. En general depende del objetivo del estudio. Así, tendremos:

• *Selección forzada*. Definido el modelo con todas las variables explicativas a estudio se inspecciona los contrastes individuales reduciendo el modelo en aquellas que no son significativas. Útil como procedimiento exploratorio

• *Regresión jerárquica*. El investigador define un orden de introducción de las variables (primero el factor de riesgo, segundo variables confundientes, tercero interacciones, etc.) y evalúa las significaciones a través de los estadísticos y contrastes de cambio. Responde a una secuencia de análisis y objetivos bien definidos

• *Regresión por 'stepwise' o etapas*. Modalidad de regresión jerárquica pero dejando que las variables entren o salgan por criterios exclusivamente estadísticos. Útil cuando buscamos el mejor modelo predictivo pero puede conducir a modelos incoherentes.

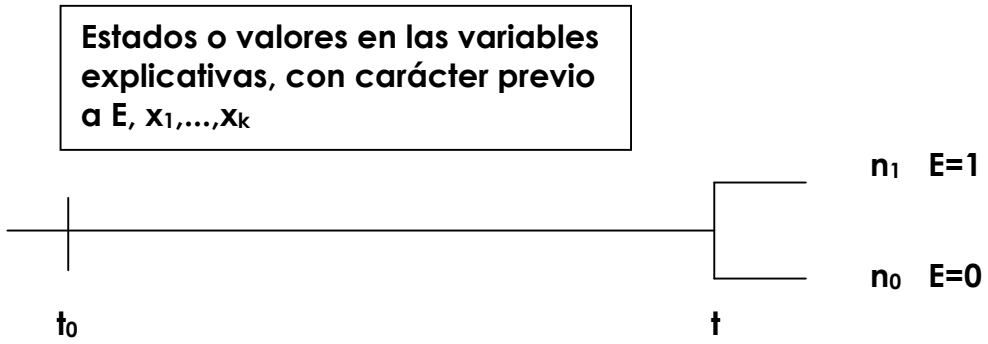
Debe tenerse en cuenta que, en general, las pruebas de hipótesis basadas en el cambio en la verosimilitud por añadir o eliminar una o más variables son más potentes y preferidas a las pruebas aisladas sobre las variables. Este criterio debe aplicarse especialmente con la introducción de interacciones en el modelo

Utilización de la regresión logística en estudios retrospectivos (casos-controles)

Notas

• Todo lo dicho hasta el momento es válido para un diseño de estudio de seguimiento de una cohorte (prospectivo). En el caso de estudios retrospectivos, la utilización del odds ratio como medida de asociación es correcta y equivalente a la de una situación prospectiva. Esto no sucede con otras medidas de asociación como el riesgo relativo.

• ¿Podemos utilizar el modelo logístico en estudios retrospectivos?. La respuesta es sí pero con limitaciones. La figura adjunta resume una situación retrospectiva



Una vez seleccionados n_1 casos (E=1) y n_0 no casos (E=0), tendremos las siguientes probabilidades:

$$\pi_1 = \frac{n_1}{N_1} = p(\text{Estar en la muestra}/E = 1)$$ **Fracción de muestreo de casos**

$$\pi_0 = \frac{n_0}{N_0} = p(\text{Estar en la muestra}/E = 0)$$ **Fracción de muestreo de no casos**

Notas

Si denotamos por:

$p'(x) =$	Probabilidad de E=1 en la muestra =	$p(E = 1/Muestra)$
$1 - p'(x) =$	Probabiilidad de E=0 en la muestra =	$p(E = 0/Muestra)$
$p(x) =$	Probabilidad de E=1 verdadera =	$p(E = 1)$
$1 - p(x) =$	Probabiilidad de E=0 verdadera =	$p(E = 0)$

A través del teorema de Bayes tendremos lo siguiente:

$$p'(x) = \frac{p(Muestra/E = 1) \cdot p(E = 1)}{p(Muestra)} = \frac{\pi_1 \cdot p(x)}{\pi_1 \cdot p(x) + \pi_0 \cdot (1 - p(x))}$$

$$1 - p'(x) = \frac{p(Muestra/E = 0) \cdot p(E = 0)}{p(Muestra)} = \frac{\pi_0 \cdot (1 - p(x))}{\pi_1 \cdot p(x) + \pi_0 \cdot (1 - p(x))}$$

Puesto que el modelo logístico es aplicado sobre los datos de nuestra muestra, tendremos que la modelización realizada será sobre el logaritmo del odds muestral:

$$\frac{p'(x)}{1 - p'(x)} = \frac{\pi_1 \cdot p(x)}{\pi_0 \cdot (1 - p(x))} = \frac{\pi_1}{\pi_0} \cdot \frac{p(x)}{1 - p(x)} \qquad (12)$$

Teniendo en cuenta que en realidad el modelo que podemos modelizar con nuestros datos es:

Notas

$$\log \frac{p'(x)}{1-p'(x)} = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$$

y el que quisiéramos realmente modelizar (al representar la verdadera probabilidad de E=1), utilizando (12):

$$\log \frac{p(x)}{1-p(x)} = \log \frac{p'(x)}{1-p'(x)} - \log \frac{\pi_1}{\pi_0} = (\beta_0 - \log \frac{\pi_1}{\pi_0}) + \beta_1 x_1 + \ldots + \beta_k x_k$$

observando que el modelo que realmente ajustamos (p'(x)) y el que quisiéramos ajustar (p(x)) difieren sólo en la constante del modelo.

• Los efectos y asociaciones detectados y cuantificados al utilizar regresión logística en situaciones retrospectivas son válidos. El modelo no puede ser utilizado en predicciones salvo que conozcamos las fracciones de muestreo de casos y controles o su relación

El MODELO DE REGRESION DE POISSON

¿Recuerda qué era la función de verosimilitud? Defínala de nuevo

• La terminología regresión de Poisson alude al tipo de función de verosimilitud de los datos. En esta modelización se supondrá que nuestras observaciones se rigen por un modelo de Poisson:

$$p(x = k) = \frac{e^{-\mu}\mu^k}{k!} \qquad k = 0,1,2,3,\ldots,\infty$$

$$E(x) = \mu \qquad D(x) = \sqrt{\mu}$$

• El modelo de probabilidad de Poisson ha sido ampliamente utilizado en el entorno sanitario debido a que es utilizado para la modelización de la ocurrencia de resultados de fenómenos poco frecuentes. Esto lo hace especialmente útil en el caso de la mortalidad y la morbilidad general o por causas específicas

• En general, si en el seguimiento de una cohorte definimos la tasa de ocurrencia (densidad de incidencia) de un fenómeno como

$\lambda = \frac{\mu}{L}$ con μ = nº de ocurrencias de un evento a estudio (defunciones)

L = Tiempo total de seguimiento de los sujetos hasta el final del estudio o la ocurrencia del evento

tendremos:

$$p(x = k) = \frac{e^{-(\lambda L)}(\lambda L)^k}{k!} \qquad k = 0,1,2,3,\ldots,\infty$$

La modelización de la regresión de Poisson. Ejemplificación con variables categóricas.

Las tasas del ejemplo seguido, ¿son densidades de incidencia? ¿Qué clase de tasas son? ¿Cómo suele calcularse la población a riesgo en estos estudios?

• Suponga que disponemos de observaciones de las defunciones y la población a riesgo de dos áreas (área 0, área 1, municipios, distritos, etc.), por grupos de edad (2 grupos, 0 '<=30 años', 1 '>30 años') (Vea los datos en la tabla 8, más adelante). Dispondremos de las siguientes tabulaciones posibles:

Tabla de defunciones por edad y área

	Area 0	Area 1
Edad 0	d_{00}	d_{01}
Edad 1	d_{10}	d_{11}

	Area 0	Area 1
Edad 0	42	122
Edad 1	300	1303

Tabla de población a riesgo (personas-tiempo) por edad y área

	Area 0	Area 1
Edad 0	L_{00}	L_{01}
Edad 1	L_{10}	L_{11}

	Area 0	Area 1
Edad 0	49300	75700
Edad 1	79300	164500

Tabla de tasas específicas por edad y área

	Area 0	Area 1
Edad 0	λ_{00}	λ_{01}
Edad 1	λ_{10}	λ_{11}

	Area 0	Area 1
Edad 0	0,00085	0,00161
Edad 1	0,00378	0,00792

El modelo de regresión de Poisson

De acuerdo con la modelización de las tasas, deduzca la expresión para los riesgos relativos ajustados (cociente de tasas) de edad=1 vs. edad=0 y de área=1 vs. área=0

• La modelización de regresión de Poisson consiste en establecer un modelo de efectos lineales de la edad y el área sobre el logaritmo de la tasa del subgrupo correspondiente:

$$\log \lambda_{ij} = \beta_0 + \beta_1(\text{Edad} = i) + \beta_2(\text{Area} = j) \qquad i = 0,1 \quad j = 0,1$$

o, equivalentemente

$$\lambda_{ij} = e^{\beta_0+\beta_1(\text{Edad}=i)+\beta_2(\text{Area}=j)} \qquad i = 0,1 \quad j = 0,1$$

• Si expresamos las tasas según este modelo, tendremos:

	Area 0	Area 1
Edad 0	$\lambda_{00} = e^{\beta_0}$	$\lambda_{01} = e^{\beta_0+\beta_2}$
Edad 1	$\lambda_{10} = e^{\beta_0+\beta_1}$	$\lambda_{11} = e^{\beta_0+\beta_1+\beta_2}$

• Con la modelización propuesta, podemos observar que el riesgo relativo de muerte para la variable edad (edad=1 vs. edad=0), como cociente de tasas, toma la misma expresión en cada área:

$$RR_{edad=1/edad=0} = e^{\beta_1}$$

• LO mismo sucede si observamos el riesgo relativo para la variable área (área=1 vs área=0), y será el mismo para cada edad:

$$RR_{área=1/área=0} = e^{\beta_2}$$

• A partir de la modelización propuesta, podemos concluir lo siguiente:

- **Los parámetros del modelo son interpretables a través de potencias de base e como riesgos relativos (cocientes de tasas)**
- **El modelo propuesto supone que el efecto de una variable explicativa es constante en los niveles de la otra (no hay interacción)**

Sobre los datos y tasas reales del ejemplo, calcule los riesgos relativos para la variable edad en cada nivel de área y para la variable área en cada nivel de edad. ¿Son constantes?

Interacciones. El modelo saturado

• Si queremos expresar un modelo que incluya la posibilidad de interacción, es decir, que los riesgos relativos para una variable no son constantes en los niveles de la otra, podemos hacerlo añadiendo un término para el producto:

$$\log \lambda_{ij} = \beta_0 + \beta_1(\text{Edad} = i) + \beta_2(\text{Area} = j) + \beta_{12}(\text{Edad} \cdot \text{Area})$$

o, equivalentemente

$$\lambda_{ij} = e^{\beta_0 + \beta_1(\text{Edad}=i) + \beta_2(\text{Area}=j) + \beta_{12}(\text{Edad}\cdot\text{Area})}$$

con esta formulación tendremos:

	Area 0	Area 1
Edad 0	$\lambda_{00} = e^{\beta_0}$	$\lambda_{01} = e^{\beta_0+\beta_2}$
Edad 1	$\lambda_{10} = e^{\beta_0+\beta_1}$	$\lambda_{11} = e^{\beta_0+\beta_1+\beta_2+\beta_{12}}$

¿Cuáles son ahora las expresiones para los riesgos relativos (cociente de tasas) de edad=1 vs. edad=0 y de área=1 vs. área=0? ¿Son constantes?

• Con la modelización propuesta, podemos observar que el riesgo relativo de muerte para la variable edad (edad=1 vs. edad=0), como cociente de tasas, no toma la misma expresión en cada área:

$$RR_{edad=1/edad=0} = e^{\beta_1} \quad \text{en área = 0}$$

$$RR_{edad=1/edad=0} = e^{\beta_1+\beta_{12}} \quad \text{en área = 1}$$

• Lo mismo sucede si observamos el riesgo relativo para la variable área (área=1 vs área=0), que tampoco será el mismo para cada edad:

$$RR_{área=1/área=0} = e^{\beta_2} \quad \text{en edad = 0}$$

$$RR_{área=1/área=0} = e^{\beta_2+\beta_{12}} \quad \text{en edad = 1}$$

• Además, el número de unidades de información necesaria para ajustar este modelo y los datos necesarios son los de la tabla 8:

Tabla 8.- Datos del ejemplo de mortalidad

Area	Edad	Defun.	Población	Tasa
0	0	42	49300	0,00085
0	1	300	79300	0,00378
1	0	122	75700	0,00161
1	1	1303	164500	0,00792

¿De cuántas unidades de información se dispone? ¿Cuántos parámetros tiene el modelo con interacción? ¿Cómo se llama un modelo como éste?

• Como se observa el número de unidades de información coincide con número de parámetros, dando lugar a que el modelo de interacción sea el saturado. Esta forma de presentación de datos, agrupando los resultados en forma de numerador y denominador de la tasa según los subgrupos de las variables explicativas es frecuente cuando las variables son categóricas (tanto dicotómicas como con múltiples categorías). La consecuencia es que podemos llegar a saturar el modelo, lo cual nos enriquece la posibilidad de interpretación de las pruebas de hipótesis de bondad de ajuste y otras como después veremos.

Deduzca la expresión (13). Interprétela.

El modelo general

• Si suponemos x_1, x_2,...,x_k variables explicativas y $\lambda(x)$ = tasa específica de ocurrencia de un evento para el perfil o subgrupo de categorías o valores $x=(x_1, x_2,...,x_k)$, la modelización de regresión de Poisson toma la forma:

En general, los datos pueden ser individuales. ¿Qué representa la tasa modelizada en ese caso? ¿Podemos llegar a saturar el modelo en ese caso?

$$\log \lambda(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

o, equivalentemente

$$\lambda(x) = e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k}$$

obteniendo una interpretación de parámetros similar al caso de la regresión logística:

$$RR_{x^*/x} = e^{\sum_{i=1}^{k} \beta_i (x_i^* - x_i)} = \prod_{i=1}^{k} e^{\beta_i (x_i^* - x_i)} \qquad (13)$$

Requerimientos en el modelo de regresión de Poisson

Notas

• Como en regresión logística, las inferencias no requerirán suposición distribucional alguna más allá que la de la propia verosimilitud. Se dirá que las inferencias son asintóticas, es decir, válidas para un n suficientemente grande. Se suponen efecto multiplicativos

Construcción de un modelo de regresión de Poisson. Etapas

• Suponga el ejemplo descrito en la tabla 8

Etapa 1: Especificación de variables y modelo propuesto

Se desea averiguar si el riesgo relativo de muerte (cociente de tasas) se ve afectado por el área, el grupo de edad o ambas. Si denotamos por

$\lambda(x)$ = Tasa específica de muerte
x_1 = Area 0 o área 1
x_2 = Edad (0 '<40', 1 '>=40')

Modelo propuesto

$$\log \lambda(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Etapa 2: Estimación del modelo

A partir de los datos disponibles, con las observaciones sobre n individuos o subgrupos de las variables:

$$\{d_i, L_i, x_{1i}, x_{2i}, \cdots, x_{ki}\}_{i=1}^{n}$$

Se trata de calcular los estimadores muestrales de los parámetros del modelo:

$(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k)$ estimadores de los parámetros $(\beta_0, \beta_1, \beta_2, \ldots, \beta_k)$

El método de estimación utilizado es el de máxima verosimilitud. La forma de la verosimilutud será la de un modelo de Poisson:

$$l(x,\beta) = \prod_{i=1}^{n} p(d_i;\beta) = \prod_{i=1}^{n} \frac{[L_i \lambda_i(x)]^{d_i} e^{-L_i \lambda_i(x)}}{d_i!}$$

con

$$\lambda(x) = e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2}$$

El resultado de la estimación del modelo propuesto es:

$$\log \lambda(x) = -7{,}136 + 0{,}7277\ \text{Area} + 1{,}568\ \text{Edad}$$

Interprete las estimaciones de los parámetros del modelo ajustado. ¿Cómo podemos interpretar β_0?

• El proceso de estimación máximo verosímil permite obtener estimaciones de:

Notas

Tabla 7.- Resultados en la estimación máximo-verosímil

$\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k)$	Estimadores de los parámetros
$l(\hat{\beta})$	Valor de la función de verosimilitud para el modelo
$\log l(\hat{\beta})$	Log-verosimilitud. Valor del logaritmo de la verosimilitud
$V = \begin{pmatrix} s^2_{\hat{\beta}_0} & & \\ \cdot & \cdot & \\ \cdot & & \cdot \\ s_{\hat{\beta}_0\hat{\beta}_k} & \cdot \; \cdot & s^2_{\hat{\beta}_k} \end{pmatrix}$	Matriz simétrica de varianzas covarianzas de los estimadores de los parámetros del modelo
$s^2_{\hat{\beta}_i} = \mathrm{Var}(\hat{\beta}_i)$	Elemento i-ésimo de la diagonal de la matriz V
$s_{\hat{\beta}_i} = \text{Error Estándar}(\hat{\beta}_i)$	
$s_{\hat{\beta}_i\hat{\beta}_j} = \mathrm{Cov}(\hat{\beta}_i, \hat{\beta}_j)$	Covarianza entre los estimadores de los parámetros

Etapa 3: Bondad de ajuste del modelo

• La bondad de ajuste del modelo se evaluará a través de la log-verosimilitud, concretamente a través de la función:

$$-2\log l(\hat{\beta}) = -2\log \text{verosimilitud}$$

siguiendo el mismo esquema jerárquico de evaluación que en el caso de la regresión logística (ver (7)).

• Sin embargo, en el caso de la regresión de Poisson podemos distinguir dos casos:

Caso 1: Disponemos de tantas observaciones como subgrupos se puedan generar a partir de las categorías de las variables explicativas, que son todas categóricas. Es el caso del ejemplo, en el que tenemos 4 observaciones, tantas como las 2x2 celdas que se pueden generar con el cruce de área x edad. En este caso, el modelo con interacción tendrá tantos parámetros como observaciones y su discrepancia será 0, al coincidir con el modelo saturado.

Caso 2: Las observaciones son sujetos individuales y/o incluyen variables cuantitativas. En este caso el modelo con interacción no suele coincidir con el saturado. Hay más observaciones que parámetros y estamos en un tipo de evaluación de bondad de ajuste que es un calco de la de regresión logística

A continuación se presenta el valor de la prueba del cociente de verosimilitudes de cada modelo vs el anterior (se van añadiendo variables, el modelo inicial es el modelo con una constante) Interprete los resultados.

Modelo	-2logveros.	p
Area	206,31	<0,001
Area Edad	550,87	<0,001
Area Edad Area*Edad	0,28	0,596

¿Cuál es la discrepancia desde el modelo con una constante hasta el saturado?

Etapa 4.- Inferencias con el modelo

• A partir de los elementos que nos produce la estimación máximo-verosímil (ver tabla 7) podemos realizar las siguientes inferencias:

• Inferencias sobre los parámetros del modelo

Asintóticamente (es decir con n grande) podemo utilizar:

Prueba de hipótesis para comprobar la significación de cada variable:

$$H_0 : \beta_i = 0 \qquad H_a : \beta_i \neq 0 \qquad i = 1,\ldots,k$$

resuelta a través del estadístico:

$$z = \frac{\hat{\beta}_i}{s_{\hat{\beta}_i}} \quad \text{(normal bajo } H_0\text{)}$$

Intervalo de confianza para cada β_i

$$I_{1-\alpha}(\beta_i) = \left[\hat{\beta}_i \pm z_{1-\alpha/2} s_{\hat{\beta}_i}\right]$$

con $z_{1-\alpha/2}$ coeficiente de una normal

La tabla adjunta muestra los estimadores y sus errores estándar para el modelo con efectos de área y edad:

Variable	$\hat{\beta}$	$s_{\hat{\beta}}$
Area	0,728	0,060
Edad	1,578	0,082

Construya pruebas de hipótesis e intervalos de confianza para los parámetros del modelo.

¿Cuánto valen los intervalos de confianza de nivel 95% para los riesgos relativos de las variables área y edad?

Notas

• Intervalo de confianza para una combinación lineal de parámetros del modelo

Dada una combinación lineal de los parámetros del modelo, con $\{a_i\}_{i=1}^{k}$ constantes conocidas:

$$C = \sum_{i=1}^{k} a_i \beta_i$$

El intervalo de confianza para C puede ser obtenido a través de la expresión:

$$I_{1-\alpha}(C) = \left[\hat{C} \pm z_{1-\alpha/2} \sqrt{Var(\hat{C})} \right] \qquad (10)$$

con:

$\hat{C} = \sum_{i=1}^{k} a_i \hat{\beta}_i$ estimador de C, combinación lineal de estimadores de los parámetros del modelo

$Var(\hat{C}) = \sum_{i=1}^{k} a_i^2 Var(\hat{\beta}_i) + 2 \sum_{i<j} \sum_{j \neq i} a_i a_j Cov(\hat{\beta}_i, \hat{\beta}_j)$ (11) varianza estimada

con varianzas y covarianzas obtenidas de la matriz de varianzas-covarianzas (tabla 7)

• Intervalo de confianza para riesgos relativos

De acuerdo con la expresión general (13) para el riesgo relativo del ocurrencia del evento a estudio entre dos subgrupos o perfiles:

$$RR_{x^*/x} = e^{\sum_{i=1}^{k} \beta_i (x_i^* - x_i)} = \prod_{i=1}^{k} e^{\beta_i (x_i^* - x_i)} \qquad (13)$$

El intervalo de confianza para el riesgo relativo puede ser obtenido como:

$$I_{1-\alpha}(RR_{x^*/x}) = I_{1-\alpha}(e^{C}) = \left[e^{\hat{C} - z_{1-\alpha/2}\sqrt{Var(\hat{C})}} ; e^{\hat{C} + z_{1-\alpha/2}\sqrt{Var(\hat{C})}} \right]$$

siendo $\hat{C} = \sum_{i=1}^{k} \hat{\beta}_i (x_i^* - x_i)$ combinación lineal de estimadores de los parámetros del modelo, y $Var(\hat{C})$ obtenida como en (11)

Sabiendo que la matriz de varianzas-covarianzas de los estimadores de los parámetros de los efectos área y edad es:

	Area	Edad
Area	0,00363	
Edad	-0,000183	0,00673

Construya el intervalo de confianza para el riesgo relativo de muerte del subgrupo area 1, edad 1 vs area 0, edad 0.

El MODELO DE REGRESION DE COX

Notas

• La regresión de Cox es conocida también como modelo de riesgos proporcionales de Cox. El contexto para su aplicación es el de una cohorte seguida en el tiempo, con observación de aparición de un evento (muerte, recidiva, enfermedad, etc.) sobre los individuos a estudio. El objetivo reside en averiguar si la mayor o menor frecuencia de aparición del evento es explicada, a lo largo del tiempo, por una colección de variables explicativas.

Algunos conceptos del análisis de la supervivencia

• A partir de n individuos a seguimiento se define:

Origen (O): Tiempo 0 o inicio del seguimiento

Evento (E): Suceso del cual esperamos su ocurrencia

Tiempo (T): Tiempo transcurrido desde el origen hasta la ocurrencia del evento o la finalización del seguimiento. No tiene porqué ser el mismo para todos los individuos

Caso Censurado: Individuo para el que al finalizar el estudio no se ha presentado el evento, bien porque no todos los individuos tienen porqué presentarlo o bien porque el estudio finaliza antes de que se presente

Tasa de riesgo o peligro: Representa la probabilidad instantánea (en un intervalo muy pequeño de tiempo) de ocurrencia del evento

$$\lambda(t) = \lim_{\Delta t \to 0} p(\text{evento en } t + \Delta t \,/\, \text{no se ha producido hasta } t)$$

Función de Supervivencia: Representa la probabilidad de que el evento se produzca después de un instante t

$$S(t) = p(\text{evento se produzca después de } t) = p(T > t) = 1 - p(T \leq t) = 1 - F(t)$$

con F(t) la función de distribución de la variable T

Tasa de riesgo acumulada: Representa la función del riesgo acumulado hasta un instante t

$$H(t) = \int_0^t \lambda(u)du$$

y presenta relación directa con la supervivencia:

$$S(t) = e^{-H(t)} \qquad H(t) = -\log(S(t))$$

Notas

Un ejemplo introductorio

• Suponga los datos de un estudio de seguimiento de la tabla adjunta

Tabla 9.- Datos ejemplo de supervivencia

Sujeto	Tiempo (T, días)	Evento (E, 1 Si 0 No)	Grupo (0, 1)
1	10	0	0
2	5	0	0
3	5	1	0
4	4	0	0
5	2	1	0
6	10	0	1
7	5	1	1
8	5	1	1
9	2	1	1
10	1	1	1

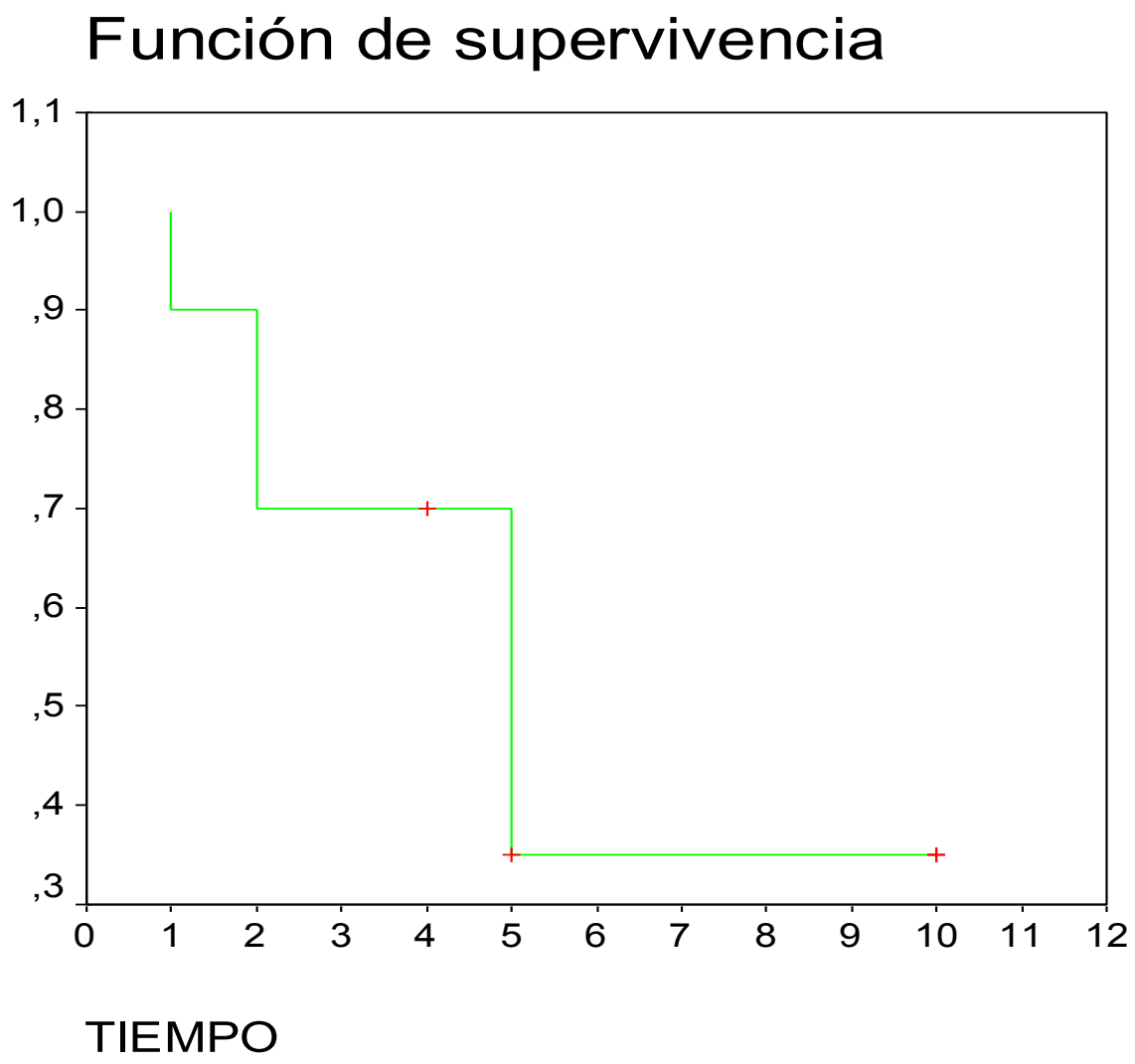
• Podemos construir la tabla de elementos de la supervivencia:

Tabla 10.- Elemento del análisis de la supervivencia

Tiempo	Sujetos a riesgo r(t)	E=1 m(t)	Tasa de riesgo $\lambda(t)=m(t)/r(t)$	$1-\lambda(t)$	S(t+1)	H(t+1)
1	10	1	1/10=0,10	0,90	0,90	0,11
2	9	2	2/9=0,22	0,78	0,70	0,35
3	7	0	0	1,00	0,70	0,35
4	7	0	0	1,00	0,70	0,35
5	6	3	3/6=0,50	0,50	0,35	1,05
6	2	0	0	1,00	0,35	1,05
7	2	0	0	1,00	0,35	1,05
8	2	0	0	1,00	0,35	1,05
9	2	0	0	1,00	0,35	1,05
10	2	0	0	1,00	0,35	1,05

Revise la construcción de la tabla e identifique los elementos estimados. Las tasas de riesgo o peligro ¿en qué intervalos están siendo estimadas??

Identifique en el gráfico adjunto los elementos de la supervivencia



El modelo de regresión de Cox. Caso de una variable explicativa

• A partir del ejemplo anterior hemos identificado y ejemplificado los elementos para el análisis de la supervivencia. La pregunta que surge ahora es ¿podemos decir que la supervivencia es diferente según la variable grupo?

Una forma de contestar esta pregunta es establecer un modelo con variable respuesta la supervivencia, concretamente la tasa de riesgo o peligro y variable explicativa el grupo. Si denotamos por x_1=grupo=(0,1), tendremos el modelo de regresión de Cox:

$$\log \lambda(x,t) = \lambda_0(t) + \beta_1 x_1$$

o equivalentemente

$$\lambda(x,t) = e^{\lambda_0(t)+\beta_1 x_1}$$

con

$\lambda(x,t)$ = Tasa de riesgo según la variable x_1, para t=0,...,∞

$\lambda_0(t)$ = Tasa de riesgo basal para x_1=0, t=0,...,∞

Como se observa en la tabla 10, a partir de las tasas de riesgo es posible construir las funciones de supervivencia, o de riesgo acumulado, por tanto el modelo podrá ser utilizado también con esta finalidad

Considerando que x_1=0 o 1, deduzca la interpretación del parámetro β_1
Discuta la interpretación de la *constante* del modelo

• Teniendo en cuenta que $x_1=(0,1)$, tendremos

Para $x_1=0$ $$\lambda(x_1 = 0, t) = e^{\lambda_0(t)}$$

Para $x_1=1$ $$\lambda(x_1 = 1, t) = e^{\lambda_0(t)+\beta_1}$$

Obteniendo así que el riesgo relativo, como cociente de tasas de riesgo, será:

$$RR_{x_1=1/x_1=0} = \frac{\lambda(x_1 = 1, t)}{\lambda(x_1 = 0, t)} = e^{\beta_1} \qquad \forall t$$

pudiendo concluir lo siguiente:

- Los parámetros del modelo son tranformables a través de la función exponencial en riesgos relativos (cocientes de tasas de riesgo)

- Los riesgos relativos así obtenidos no dependen del tiempo, son constantes para cualquier momento t. Este resultado es consecuencia de la estructura del modelo. Es conocido como hipótesis de **proporcionalidad de riesgos** dando así nombre al modelo, que establece que sea cual sea el instante t al que nos refiramos, la tasa de riesgo en uno de los grupos es proporcional a la del otro grupo. El factor de proporcionalidad es el riesgo relativo

Discuta la proporcionalidad de riesgos. ¿Cree que es una suposición aceptable de forma general?

Notas

• Una forma de detectar y estimar la desviación de la hipótesis de proporcionalidad de riesgos puede ser introducir en el modelo un término que incorpore el tiempo:

$$\lambda(x,t) = e^{\lambda_0(t)+\beta_1 x_1+\varphi(x_1 \cdot t)}$$

siendo t el tiempo de seguimiento

Con este término de interacción podemos capturar la falta de proporcionalidad de riesgos, teniendo que:

Para $x_1=0$ $$\lambda(x_1 = 0,t) = e^{\lambda_0(t)}$$

Para $x_1=1$ $$\lambda(x_1 = 1,t) = e^{\lambda_0(t)+\beta_1+\varphi t}$$

Obteniendo así que el riesgo relativo será función del tiempo:

$$RR_{x_1=1/x_1=0}(t) = \frac{\lambda(x_1 = 1,t)}{\lambda(x_1 = 0,t)} = e^{\beta_1+\varphi t} \qquad t = 0,\ldots,\infty$$

Con esta formulación se puede estimar el riesgo relativo aunque dependa del tiempo, limitados a la forma de evolución exponencial

Generalización a múltiples variables explicativas

Deduzca la expresión (13). Interprétela.

• Si suponemos x_1, x_2,...,x_k variables explicativas y $\lambda(x,t)$ = tasa de riesgo en t de ocurrencia de un evento para el perfil o subgrupo de categorías o valores $x=(x_1, x_2,...,x_k)$, la modelización de regresión de Cox toma la forma:

$$\log \lambda(x,t) = \lambda_0(t) + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k$$

o, equivalentemente

$$\lambda(x,t) = e^{\lambda_0(t)+\beta_1 x_1+\beta_2 x_2+...+\beta_k x_k}$$

obteniendo una interpretación de parámetros similar al caso de los modelos precdentes:

En regresión de Cox, los datos son siempre individuales. ¿Podemos llegar a saturar el modelo en ese caso?

$$RR_{x^*/x} = \frac{\lambda(x^*,t)}{\lambda(x,t)} = e^{\sum_{i=1}^{k}\beta_i(x_i^*-x_i)} = \prod_{i=1}^{k} e^{\beta_i(x_i^*-x_i)} \qquad (14)$$

Requerimientos en el modelo de regresión de Cox

• Como en los modelos precedentes, las inferencias no requerirán suposición distribucional alguna más allá que la de la propia verosimilitud. Se dirá que las inferencias son asintóticas, es decir, válidas para un n suficientemente grande. Se suponen efectos multiplicativos

Construcción de un modelo de regresión de Cox. Etapas

Observe los datos para el ejemplo en el anexo 3

• Suponga como ejemplo observaciones sobre 60 individuos trasplantados, con las siguientes variables:

TIEMPO: Tiempo de seguimiento (días) EDAD: En años
EVENTO: Resultado al final del seguimiento 0 'Favorable' 1 'Desfavorable'
TRATA: 0 'No Tratamiento' 1 'Tratamiento SEXO: 1 'Hombre' 2 'Mujer'
SCORE: Puntuación de compatibilidad entre tejidos

<u>Etapa 1</u>: Especificación de variables y modelo propuesto

Con el modelo propuesto, ¿cómo interpretamos $\lambda_0(t)$

Se desea averiguar si el tratamiento disminuye el riesgo (la tasa de riesgo) de resultado desfavorable (p.ej. muerte) controlando (ajustando) por la puntuación de compatibilidad entre tejidos.

Si denotamos por

$\lambda(x,t)$ = Tasa instantánea de riesgo
x_1 = Tratamiento 0 'No' 1 'Si'
x_2 = *score* . Variable continua

Modelo propuesto

$$\log \lambda(x,t) = \lambda_0(t) + \beta_1 x_1 + \beta_2 x_2$$

Etapa 2: Estimación del modelo

Notas

A partir de los datos disponibles, con las observaciones sobre n individuos o subgrupos de las variables:

$$\{E_i, T_i, x_{1i}, x_{2i}, \cdots, x_{ki}\}_{i=1}^{n}$$

E_i= Resultado del evento en el sujeto i
T_i= Tiempo de seguimiento del sujeto i

Se trata de calcular los estimadores muestrales de los parámetros del modelo:

$(\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k)$ estimadores de los parámetros $(\beta_1, \beta_2, \ldots, \beta_k)$

El método de estimación utilizado es el de máxima verosimilitud

- Supuestos ordenados los instantes de ocurrencia del evento sobre los sujetos, digamos:

$$t_1 < t_2 < \ldots < t_n \qquad t_i \neq t_j \quad \forall i, j$$

- Dado R_i conjunto de individuos que sobreviven a t_i (su momento de ocurrencia del evento es >= t_i

- La función de verosimilitud se obtiene a partir de la probabilidad condicional de que cada uno de los n individuos presente el evento E en t_i dado que exactamente uno de los R_i lo presenta en t_i

Interprete los valores obtenidos para los estimadores de los parámetros de tratamiento y score

$$l(x,\beta) = \frac{\text{Probabilidad de evento en } t_i}{\text{Probabilidad de evento en otro } t_j \geq t_i} =$$

$$\prod_{i=1}^{n}\left[\frac{\lambda(x,t_i)}{\sum_{j\in R_i}\lambda(x,t_j)}\right] = \prod_{i=1}^{n}\left[\frac{e^{\beta_1 x_{1i}+\ldots+\beta_k x_{ki}}}{\sum_{j\in R_i} e^{\beta_1 x_{1j}+\ldots+\beta_k x_{kj}}}\right]$$

Nótese que la verosimilitud no permite estimar $\lambda_0(t)$ que será estimada por métodos no paramétricos

El resultado de los estimadores de los parámetros del modelo propuesto es:

Variables en la ecuación

	B	Exp(B)
TRATA	,346	1,413
SCORE	1,456	4,288

• El proceso de estimación permite obtener también los elementos descritos en la tabla 7 (varianzas, covarianzas y logverosimilitudes)

Etapa 3: Bondad de ajuste del modelo

• La bondad de ajuste del modelo se evaluará a través de la log-verosimilitud siguiendo el mismo esquema jerárquico de evaluación que en el caso de los modelos precedentes (ver (7))

Etapa 4.- Inferencias con el modelo

• A partir de los elementos que nos produce la estimación máximo-verosímil (ver tabla 7) podemos realizar las siguientes inferencias:

• Inferencias sobre los parámetros del modelo

Asintóticamente (es decir con n grande) podemo utilizar:

Prueba de hipótesis para comprobar la significación de cada variable:

$$H_0 : \beta_i = 0 \qquad H_a : \beta_i \neq 0 \qquad i = 1,\ldots,k$$

resuelta a través del estadístico: $z = \dfrac{\hat{\beta}_i}{s_{\hat{\beta}_i}}$ (normal bajo H_0) o su cuadrado, test de Wald

Intervalo de confianza para cada β_i

$$I_{1-\alpha}(\beta_i) = \left[\hat{\beta}_i \pm z_{1-\alpha/2} s_{\hat{\beta}_i}\right]$$ con $z_{1-\alpha/2}$ coeficiente de una normal

A continuación se presenta el valor de la discrepancia de cada modelo vs el saturado (se van añadiendo variables). Los valores p se refieren a la significación en la disminución de la discrepancia de uno a otro modelo.

Modelo	-2logveros.	p
Inicial	219,95	
Trata	216,18	<0,052
Trata Score	195,75	<0,001
Trata Score Trata*Score	193,76	0,159

¿Cuánto vale el test del cociente de verosimilitudes para cada modelo vs el anterior?¿Cuántos grados de libertad tiene cada prueba? ¿Cuántos grados de libertad tiene la prueba que evaluaría si la discrepancia de cada modelo es significativa respecto del modelo saturado? Interprete los resultados

Algunas cuestiones adicionales

Confusiones e interacciones

Son tratadas como en los modelos precedentes. Las tablas adjuntas muestran resultados de los modelos con trata, trata y score y trata, score e interacción entre ambos. Discuta la confusión en el efecto de trata.

Variables en la ecuación

	B	ET	Wald	gl	Sig.	Exp(B)	95,0% IC para Exp(B)	
							Inferior	Superior
TRATA	,707	,367	3,710	1	,054	2,027	,988	4,161

Variables en la ecuación

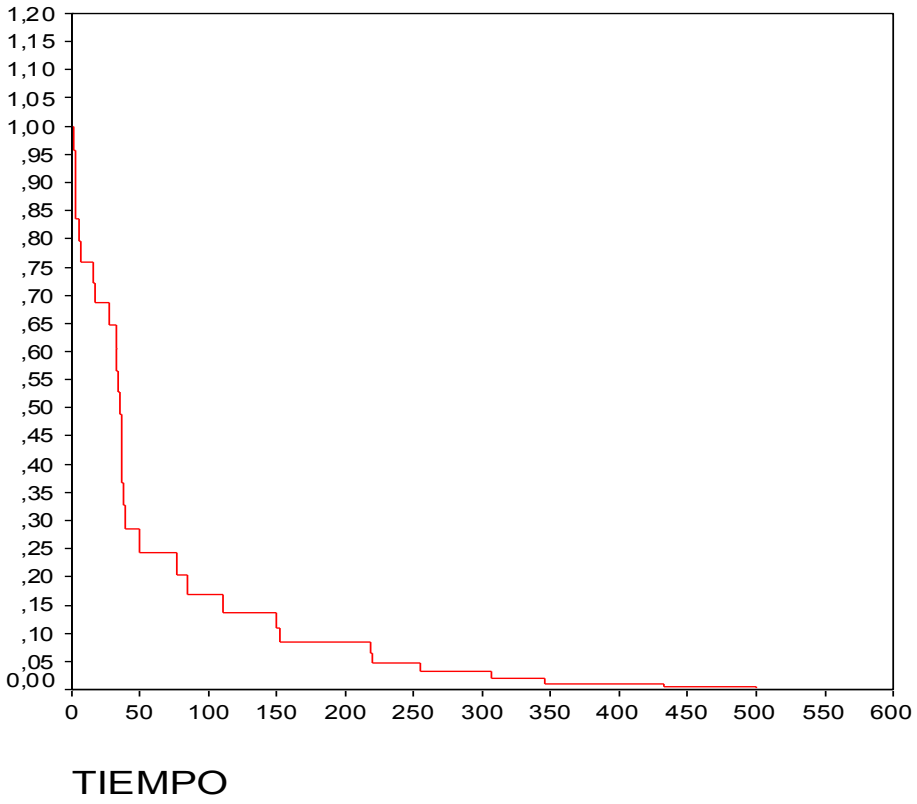
	B	ET	Wald	gl	Sig.	Exp(B)	95,0% IC para Exp(B)	
							Inferior	Superior
TRATA	,346	,387	,800	1	,371	1,413	,662	3,017
SCORE	1,456	,318	20,985	1	,000	4,288	2,300	7,994

Variables en la ecuación

	B	ET	Wald	gl	Sig.	Exp(B)	95,0% IC para Exp(B)	
							Inferior	Superior
TRATA	1,426	,874	2,660	1	,103	4,162	,750	23,089
SCORE	2,091	,548	14,544	1	,000	8,091	2,763	23,695
SCORE*TRATA	-,848	,599	2,001	1	,157	,428	,132	1,387

Gráficos de supervivencia acumulada

A partir del modelo ajustado es posible obtener gráficos para la supervivencia acumulada estimada en diferentes perfiles de las variables explicativas. El gráfico adjunto muestra esta función para trata=0 (no tratar) y score=2:



Estime aproximadamente la probabilidad acumulada de supervivencia a los 100 días del seguimiento para sujetos no tratados y score=2

Intervalos de confianza para riesgos relativos entre perfiles distintos

Se obtienen igual que en los modelos anteriores (ver regresión de Poisson, pag. 80)

BIBLIOGRAFIA

1. Anderson S, Auquier A, Hauck WW, Oakes D, Vandaele W, Weisberg HI. Statistial Methods for comparative studies. New York: Wiley & Sons, 1980

2. Breslow NE, Day NE. Statistical methods in cancer research. Volume I: The analysis of case-control studies. Lyon: IARC, 1980

3. Breslow NE, Day NE. Statistical methods in cancer research. Volume II: The design and analysis of cohort studies. Lyon: IARC, 1987

4. Hosmer DW, Lemeshow S. Applied logistic regression. New York: Wiley & Sons, 1989

5. Kleinbaum DG, Kupper LL, Morgenstern H. Epidemiologic research: Principles and quantitative methods. Belmont: LLP, 1982

6. Kleinbaum DG, Kupper LL, Muller KE. Applied regression analysis and other multivariate methods. Boston: PSW-Kent, 1988

7. Mahesh KB, Machin D. Survival analysis: A practical approach. New York: Wiley & Sons, 1995

8. Neter J, Wasserman W, Whitmore GA. Applied statistics. Boston: Allyn & Bacon, 1993

9. Peña, D. Estadística: Modelos y métodos. 2. Modelos lineales y series temporales. Madrid: Alianza, 1987

10. Schlesselman JJ. Case-control studies: Design, conduct, analysis. New York: Oxford University Press, 1982